MSc. Thesis
Computer Engineering

2nd of June 2020

Aarhus University
Department of Engineering

Clinical Decision Support for Physiotherapy based Rehabilitation

Tenna Boeriis Rasmussen (201406382)
Cecilie Østergaard Moriat (201405949)

Supervised by Christian Fischer Pedersen

Preface

This master's thesis has been accomplished at the Department of Engineering at Aarhus University to fulfill the requirements for acquiring the Master of Science degree in Computer Engineering.

We would especially like to thank our supervisor, Christian Fischer Pedersen for all guidance and advice during the thesis work.

Furthermore, we would like to thank Aalborg Municipality and DigiRehab for facilitating the data used throughout this project. We are thankful for the insight and domain knowledge we have obtained in this thesis. Additionally, thanks to Michael Harbo, DigiRehab, for taking the time to thouroughly introduce us to the vast field of physiotherapy based rehabilitation.

Aarhus, June 2nd, 2020

Cecilie Østergaard Moriat

Tenna Boeriis Rasmussen

Abstract

Today, community-dwelling citizens receiving home care experience a gradual decline in physical capacity and receive rehabilitation to counter this. However, the referral process differs across municipalities and uncertainties in the clinical judgement of rehabilitation referrals pose a challenge. These challenges are examined in this thesis, which is conducted in collaboration with Aalborg Municipality and DigiRehab in order to provide a baseline for the further studies on the KL signature project *Intelligent rehabilitation and targeted public assistance for citizens*.

This thesis combined data regarding the physiotherapy based rehabilitation programmes of citizens with data regarding their loans of assistive technologies. Two objectives for a citizen's benefit from physiotherapy-based rehabilitation were defined and examined. An extensive work of preparing the raw data was conducted in order to fit these objectives. Furthermore, classification models based on scientifically proven machine learning algorithms have been designed, implemented, tested and documented adhering to best practices including the use of cross-validation and AUC metrics.

This thesis found that for all experiments, the use of assistive technology information improved on the prediction performance. The models were optimised in terms of feature selection to increase prediction performance. Prediction capabilities were finally assessed on an independent validation set to ensure a trustworthy measurement for the model's generalisation performance, yielding prediction AUCs similar to comparable studies. With the use of LIME, illustrative explanations for a local approximation of the prediction of the citizen's rehabilitation potential were created. These were compared to the intrinsically available information of the models to examine their strengths and drawbacks. This provided a great insight into the model and its usage of the predictors.

Resumé

En gradvis forværring af den fysiske formåen ses hos borgere, der modtager hjemmepleje. For at modvirke dette tildeles rehabiliteringsforløb. Der forekommer imidlertid store forskelle i tildelingen af rehabilitering på tværs af kommuner. Dette, samt usikkerheder i den kliniske vurdering ved visitation til rehabilitering, er udfordringer for sundhedssystemet. I dette speciale undersøges disse udfordringer i et samarbejde med Aalborg kommune og DigiRehab for at danne et solidt grundlag for videre undersøgelser i et af Kommunernes Landsforenings signaturprojekter, *Intelligent rehabilitering og målrettet tilbud til borgere*.

Dette speciale kombinerer data fra borgeres fysioterapi-baserede rehabiliteringsforløb med data der beskriver deres lån af hjælpemidler. For at undersøge, hvordan borgerne får gavn af et rehabiliteringsforløb, er der udarbejdet to definitioner, som beskriver, hvordan dette kan evalueres. Et særligt arbejde er lagt i at undersøge og forberede dataen for at tilpasse den til de to målsætninger. Derudover er videnskabeligt funderede klassifikationsalgoritmer designet, implementeret, afprøvet og dokumenteret.

Dette speciale har vist, at anvendelsen af information omhandlende hjælpemidler forbedrede AUC'en for samtlige eksperimenter.

De udviklede klassifikationsalgoritmer er optimeret ved hjælp af feature selection for at opnå en højere AUC. Ved anvendelsen af cross-validation og et separat valideringssæt, kunne modellernes evne til at generalisere vurderes på en pålidelig måde, og de resulterende AUC-værdier er sammenlignelige med lignende studier. Ved hjælp af LIME frameworket kunne algoritmernes forudsigelser ledsages af illustrative forklaringer. Disse blev sammenholdt med iboende information om modellerne for derved at opnå en insigt i underliggende mønstre i dataen.

Contents

1

Introduction

1.1 Background

During the last decade a focus on supporting and assisting elderly citizens in staying self-reliant and community dwelling has reduced the number of Danish citizens living at care homes by 6.4%. At the same period (2010 to 2019) the total number citizens aged 65 and above has increased by 25% [1].

However, ageing gradually causes frailty with reduced physical fitness, loss of muscle strength, and worsened balance [2]. This is often countered by a steadily increasing number of assistive devices along with an increase in home care. The purpose of assistive technology is to remedy the consequences of permanently reduced functional capacity and to the greatest extent ease the patients day-to-day life in order to remain self-reliant. Yet, this treatment usually results in a downward spiral for the citizen where one assistive device leads to the next, while frailty rises, along with the risk of falling [3, 4]. Conversely, multiple studies show that being physically active can diminish the effects of frailty and increase self-reliance among elderly [5].

In the light of the above, there has been a greater focus on how physical rehabilitation can support occupational therapy for patients in home care, and in January 2015, Denmark adopted § 83 a, stating that the municipalities are obliged to offer temporary, short-term rehabilitation for citizens with functional impairment if it is considered that rehabilitation will improve the functional capacity and thereby reduce the need for help [6].

Still, there seem to exist an uneven usage of this section in Denmark, as the referral procedures usually are subjectively conducted. In 2017, the Danish Center for Social Science Research, published an assessment of rehabilitation practices in Denmark [7]. The assessment describes the citizens and their rehabilitation programmes, analyses the development in their functional capability and their experience with accomplishing rehabilitation. The assessment has examined the reasons for initiations of rehabilitation programmes for two municipalities which shows that 41% of the programmes are offered in continuation of a hospitalisation while 29% of the programmes are started on the basis of an re-assessment of the citizens currently granted home help. The rest are given in continuation of a temporary or longer stay at a rehabilitation center or as a result of employees tracing citizens which are considered to obtain some benefit from a rehabilitation programme.

Even so, an analysis [8] published by the Benchmark Unit of Ministry of Social Affairs and the Interior in October 2019 based on data from 17 Danish municipalities, reports large variations in physical therapy referrals of citizens. The analysis explains these by the referral procedure, which is approached in various ways across municipalities leading to an uneven usage of rehabilitation among citizens. There might be an element of subjectivity in the decision of whether a citizen has potential for improving their functional capability, or it might be based on distinct measures. The variations might also imply differences in the standard of service among municipalities.

This lack of reliability is also supported by a study from 2000 [9], which shows uncertainty in clinical judgment for rehabilitation referrals. Thus, it is interesting and highly relevant to investigate data-driven decision support for rehabilitation referrals in a home care setting.

1.2 Motivation

Clinical decision support systems (CDSSs) based on statistical- and machine learning methods have achieved good results in various areas of health care [10–20], and it is thus interesting to investigate the potential of applying similar methods for rehabilitation in a home care setting. This has been the focus of three studies based in Canada and Taiwan [18–20]. However, none of these have investigated how information about assistive technology can be utilized in this context. Further details regarding state of the art research is presented in chapter 2.

This thesis will use data collected from the municipality of Aalborg in collaboration with DigiRehab, which includes information about assistive technology and exercise programs of citizens in home care. The data will be used to study how statistical learning can identify citizens who will benefit the most from rehabilitation.

1.3 Collaborators

In 2019, Aarhus University completed a pilot project which identified typical loan sequences of assistive devices and a preliminary research in predicting the future devices of a citizen. The collaborators were Aalborg Municipality, DigiRehab, Kommunernes Landsforening (KL), Aarhus Municipality and University College Nordjylland. DigiRehab is a company providing a digital exercise app aimed at elderly citizens with homecare assistance. This app is used by Aalborg Municipality to assist the social- and health service assistant staff in the management of physical rehabilitation services to citizens. Data for the project was gathered from this app as well as a retrieval of assistive device data from the municipality.

Regarding an agreement covering the economy for municipalities and regions in 2020, the government, KL and Danish Regions launched 15 signature projects to try out artificial intelligence in municipalities and regions [21, 22]. One of these is *Intelligent rehabilitation and targeted public assistance for citizens*. Aalborg Municipality is the project

owner and leader, and is responsible for gathering the data used throughout this project in close collaboration with DigiRehab.

The overall aim of this signature project is to use artificial intelligence to offer and target citizens with the training they are most likely to benefit from. This should be done by comparing data regarding loans of assistive technologies with data from conducted rehabilitation programmes to study the correlation of these. Additionally, the project aims to identify citizens with an elevated risk of falls to initiate fall prevention programs more efficiently.

The work for this thesis is conducted simultaneously with the start-up of the signature project. This implies cooperation with Aalborg Municipality and DigiRehab regarding the objectives and findings of the project. Therefore, this thesis will engage in work applying state of the art methods and approaches for the available data. As the signature project is expected to continue over a span of two years, the findings of and methods used throughout this thesis study can support the future work of the signature project.

1.3.1 Data access and limitations

From the start of the thesis work, rehabilitation data from DigiRehab and data of assistive technology associated with citizens from Aalborg were made available. Additionally, DigiRehab rehabilitation data of citizens from Viborg Municipality were available. However, the associated data about the assitive technology was not available for the Viborg data. Thus, it was decided to limit the study to only include the data from Aalborg.

As mentioned above, the signature project had a separate aim to identify citizens in risk of falling. The literature of this was researched and a plan for including this objective in the present thesis was devised. However, because of unrelated circumstances it was not possible to obtain the prospected data in due time. Thus, the objective of investigating fall prediction was ultimately excluded from the scope of the thesis project. Refer to appendix A for the preliminary investigations conducted in this area.

1.3.2 Identifying citizens who will benefit from exercise

When diving deeper into the objective of defining citizens who will benefit from rehabilitation it is relevant to determine what characterises *benefit*. As this project is developed in collaboration with external partners who have insights and domain knowledge, the definition of benefit has been derived with input from the partners and with consideration of the signature project objectives. Firstly, § 83 a [6], as mentioned above, ultimately aims for a reduction in need for home care help. Thus, it is decided that a reduction in need for help can be used to determine benefit of exercise. This measure is presented in definition 1.1.

Definition 1.1: Benefit based on a citizen's need for help

A citizen benefits from a physical rehabilitation programme if their need for home care help has decreased after the programme has ended.

However, another measure of benefit is also defined. As a necessity to achieve any benefit from rehabilitation, actually performing the assigned physical exercise is an invariable requirement. Thus, it is relevant to identify citizens who will complete a rehabilitation programme successfully. The assessment of rehabilitation practices in Denmark from 2017 [7] concludes that citizens having completed a rehabilitation programme obtains significant improvement in their physical capabilities and an equivalent significant improvement in their own experience of their functional capability. DigiRehab has defined a successful programme as having completed training sessions in at least eight out of 12 weeks [23]. This definition is based on their experience as state-authorised physical therapists and their experience in the field of rehabilitation. Throughout this thesis, it is defined as another measure of benefit as seen in definition 1.2.

Definition 1.2: Benefit based on a citizen's completion of a rehabilitation programme

A citizen benefits from a physical rehabilitation programme if they succeed in training in at least eight out of 12 weeks.

These two definitions of benefit are central to defining the problem formulation of the present master thesis.

1.4 Problem formulation

The focus of the present master's thesis is to design, implement, test, and document explainable machine/statistical learning algorithms for clinical decision support for physiotherapy based rehabilitation. The decision support system should ease evidence-based decision making for physiotherapists, occupational therapists, home carers, and other care givers with regard to identifying citizens that will benefit from physiotherapy based rehabilitation.

To determine the benefit, a bifold approach is taken, where the following two objectives are investigated:

- identify citizens who will achieve a beneficial development in their need for home care help resulting from physiotherapy based rehabilitation, and
- identify citizens who will complete a physiotherapy based rehabilitation program successfully.

The methods employed shall provide data analytics with high transparency to ensure that the basis for decisions is substantiated for the care givers. The algorithms must be optimized with regard to predictive performance. Moreover, the algorithms must be proven functional via concrete experiments. To quantify the quality of the algorithms, comparative evaluations should be carried out via objective metrics. As an overall assessment of the algorithms and the decision support system, comparisons with state of the art should be conducted.

2

State of the art

This chapter presents the state of the art of clinical decision support systems relevant to this thesis together with a more in-depth presentation and evaluation of scientific research regarding rehabilitation in the home care sector.

2.1 Clinical decision support systems

Within the field of health care, the development and evaluation of clinical decision support systems (CDSSs) have long been an active area of research [24, 25]. For a CDSS to aid medical personnel, the system must be able to provide accurate suggestions. Multiple studies have looked into providing systems capable of making predictions that can help determine the right treatment of patients in clinical settings using statistical- and machine learning approaches [10–20]. Valdes et al. used boosted tree to develop a model for clinical decision support of radiotherapy treatment [10], and in [11] Horng et al. applied SVM, logistic regression, naïve Bayes, and random forest to predict sepsis for in-hospial patients. By use of decision tree analysis [12] has developed an algorithm to assess the risk of first time falling for home care clients in Canada.

When looking into work with related methods a retrospective cohort study from Taiwan in 2018 [13], where the development in activities of daily living (ADL) was predicted for 365 post-stroke patients, is of interest. The methods used were logistic regression, random forest and support vector machines validated using 5-fold cross-validation and AUC as evaluation metric. As predictive input various features related to nutrition, motor function, cognition and degree of disability were available. As target value of the prediction the study used the Barthel Index (BI), which they transformed from continuous to categorical in both a binary version and a three level version, where high, low, and medium BI categories were defined. Even though the setting of the study differs, the methods applied are relevant to this thesis, as the data set is of similar small size, thus, robust evaluation methods are of interest and moreover, the conversion of a continuous target variable into a categorical variable is applied in the present thesis.

Two studies[14] and [15] both apply a moving window approach for feature creation in predictive studies for clinical decision support regarding chronic obstructive pulmonary disease. However, it seems that the non-independence from including multiple windows of data from the same test subjects has not been considered. This may

result in over-fitting to specific subjects. In a separate study[16] from 2019 with similarly structured longitudinal data of diabetes patients, this non-independence is accounted for by introducing a mixed effects model to resolve the non-independence. In the present thesis the potential non-independence from longitudinal data is also considered.

2.2 Rehabilitation in home care

When narrowing the search to predicting rehabilitation potential in a home care setting, only a few studies evaluate the effects of statistical- and machine learning-based clinical decision support. These are described the following.

In a study [17] from 2015 the Geriatric Health Systems Group of the University of Waterloo in Canada looked into identifying the client characteristics most relevant in predicting who will receive rehabilitation services. They chose to use the machine learning methods LASSO and Random Forest. However, this study does not predict who will benefit from rehabilitation, but rather who receives rehabilitation services.

Another study [18] looked into providing clinical decision support for home care patients in Taiwan with a focus on occupational therapy service referrals. In this study, two algorithms were developed to predict which clients in long term care might need and benefit from home- and community-based occupational therapy. The algorithms presented in the paper are based on logistic regression. This is also a method used in the present study. However, the predictions in the Taiwan-based study were evaluated on a basis of clinical judgement by occupational therapists prior to any therapy. As presented in the introduction of this thesis, clinical judgement may not be consistent. In the present thesis the benefit of therapy is based on actual measurements after receiving physical therapy. Furthermore, the study from Taiwan used predictors based on self-reported health conditions, which may not accurately reflect the health status of the participants. Finally, the algorithms developed in the Taiwanese study predicts who will need occupational therapy, which is a broader definition of home- and community-based service, than the specific focus on physical therapy in this study.

Other work done in the area is seen in two Canadian studies [19, 20] from 2007 on a project called "Inforehab Home Care" [26] by a research group from the University of Waterloo in Canada. They apply the machine learning methods: K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) to predict rehabilitation potential for home care clients. In addition to the developed algorithms, they also use the SVM-model to help identify important variables in the prediction of rehabilitation potential. Rehabilitation potential is defined as either improvements in Activities of Daily Living (ADL) or discharge from home care, this is similar to the reduction in need for home care measure used in the present study. However, different machine learning methods are investigated in the present study.

In both the Taiwan-based study and the Canadian project the data used sets the research apart from the present study as the models in [19, 20] are trained on Canadian home care data and the models in [18] are trained on self-reported health information

from Taiwan, whereas the present study is based on data from Danish home care citizens and the available information in the datasets, and thus the predictors used, differs. In general healthcare systems and utilization of home care services differ across different countries and regions, making it relevant to conduct a local study.

Furthermore, there are, to the knowledge of the authors of this thesis, no study which has investigated the effect of including information about assistive technology as a predictor for models determining who will benefit from enrolling in a programme of physical rehabilitation.

2.3 Summary

This thesis will leverage the knowledge of state of the art approaches from the field of related work. Many clinical decision support systems have managed to take advantage of information in available data by use of machine learning methods in clinical settings [10–20]. The knowledge about these state of the art methods can be leveraged in the present thesis. Furthermore, it was found that robust methods such as Cross-Validation and AUC are widely used for selection and assessment of models in the related work[11, 13, 19, 20].

Within research of rehabilitation potential, studies have used different measures related to the activities of daily living to predict benefit from rehabilitation. In that case they have used a threshold to transform continuous variable into a dichotomous variable to indicate *benefit* or no *benefit* [13, 19, 20].

However, studies investigating data-driven and machine learning based approaches to predict rehabilitation potential in home care generally focus on a broader term of both occupational- and physiotherapy-based rehabilitation. In the present thesis the focus will specifically be on physiotherapy-based rehabilitation. Furthermore, there is an absence in investigating how the information about assistive technology can be utilised to determine who will benefit from rehabilitation. The present thesis will look into filling this absence by applying state of the art data-driven methods inspired by the mentioned related work.

3

Machine learning in healthcare

The health care sector is constantly challenged by rising prices on medicine and advanced treatments. This combined with a growth in chronic illnesses, and the positively increased life expectancy leads to a persistent need for prioritization and streamlining to reduce health care costs [27, 28]. To help practitioners prioritise and make qualified decisions, clinical decision support systems (CDSSs) of many forms have been implemented in clinical workflows. These systems have the possibility to provide a significant improvement in practitioner performance. Especially in applications where the system is able to automatically provide the user with a recommendation at the time and place of the decision to be made [24, 25].

Furthermore, recent technological developments and rise in available data provide great opportunities for advances in the applied clinical decision support. Through the employment of machine learning methods for complex statistical analysis the accuracy of CDSS recommendations can be improved [20, 29]. This has spiked an interest in the utilisation of such methods [21, 22, 30, 31]. However, the potential increase in prediction accuracy provided by the more complex and opaque machine learning methods comes at the cost of less transparency. When applied in a sensitive field like health care, it is important to ensure that the methods do not introduce unintended errors or bias and discrimination towards specific groups in the population. In a study from the 1990s [32], in which the aim was to predict and identify high risk pneumonia patients, various models were evaluated. A less accurate but intelligible rule-based model disclosed that asthmatic patients were predicted to be more likely to survive pneumonia, even though they are known to be high risk patients. This unwanted and counter intuitive result reflected a pattern in the training data, which occurred because asthmatic patients were admitted directly to the ICU and received special care. Such errors and even more subtle bias in the training data may be hard to discover when using opaque methods. In the pneumonia example it was decided to trade accuracy for explainability and trust with the selection of a logistic regression model as opposed to the more accurate but opaque neural network model. As the present study is developed to support health care practitioners, it is important to ensure transparency of the predictions made to instill trust in the users. In the following subsection different approaches to achieve transparency for decision support systems are discussed.

3.1 Decision support and transparency

For machine learning-based systems to be utilised and widely adopted in the field of health care, it is important to acknowledge the importance of trust and transparency. Practitioners and citizens must trust the system providing recommendations to be accurate and unbiased if they are to accept the suggestions provided by it. A way of achieving trust is through explainability, where predictions are substantiated by an associated explaination as shown in figure 3.1. There is currently a big focus on being able

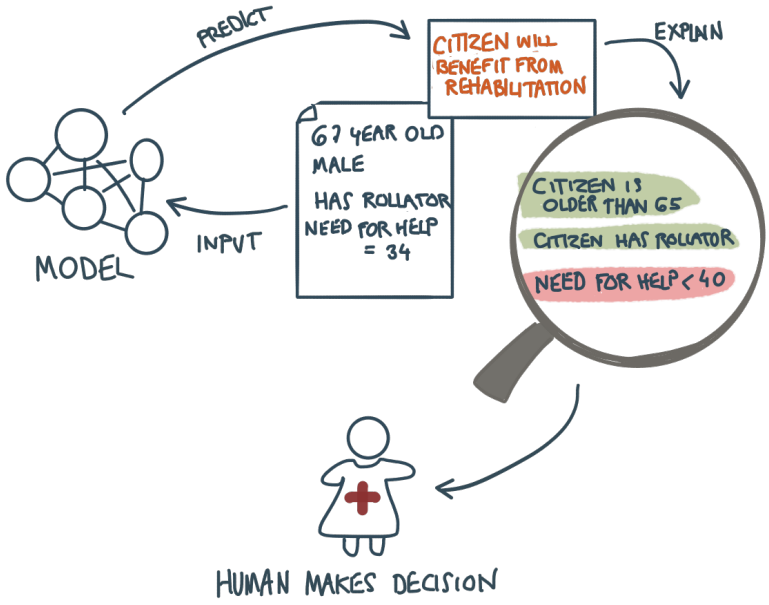


Figure 3.1: *Using explainations in decision support systems.*

to explain why a decision has been made and therein also why a model has provided a given prediction. Explainability is for example built into the European General Data Protection Regulation (GDPR), with the notion of a right to an explaination in cases of automated decision-making [33]. Explainable Artificial Intelligence (XAI) is an active field of research that works towards providing explanations for machine learning predictions [34, 35].

There are different ways of achieving explainability. One way is through the use of models that are intrinsically transparent. These are simpler models, such as logistic regression and decision trees. These are possible for humans to comprehend and explain. A different approach to achieve explainability is through approximation, where a second model accompanies the prediction model. While the model providing predictions may be opaque in nature, the second model circumvents this by approximating an explanation of the first model. This approximation can be global, covering the whole model, or it can be local providing an explanation for a specific prediction outcome.

In order to evaluate the transparency of the machine learning models developed

in the present study, the different approaches to explainability are categorized. This resulted in the following four categories:

- **Intrinsic** Achieved using simple models that are intrinsically transparent.
- **Global Approximation** Provides a global explanation of the full model.
- **Local Approximation** Provides a local explanation of a specific prediction.
- **Opaque** Using opaque models with little or no transparency, where no explanation is provided.

As this thesis is focused on providing decision support for non-technical home-care workers the methods of interest should aid in providing an explanation for specific predictions rather than the whole model. It is seen that providing too extensive information about a decision does not improve trust in the model [36]. Thus, methods that are intrinsically transparent and can provide information about feature importance are of interest along with methods that are accompanied by a second model providing local explanations of predictions.

3.1.1 Methods for local approximation

As transparency and explainability have become a focus within the machine learning world, developing methods that can provide local approximations to explain a prediction are an active area of research. Two methods for achieving this are *Local Interpretable Model-Agnostic Explanations* (LIME) [35] and *SHapley Additive exPlanations* (SHAP) [37]. The main difference is:

- **LIME** trains a local model on new random samples in the proximity of the input.
- **SHAP** uses shapley values to asses the contribution of each feature in combination with all other features

This makes SHAP computationally complex, whereas the random sampling in LIME is more efficient. Although SHAP may in be more precise, the computational complexity makes it less applicable in a field where the decision support should be integrated into the workflow and presented to the user promptly without introducing delays and blocking the workflow. For this reason the more computationally light weight method LIME is selected to be utilised in the present thesis. To investigate the potential of LIME, it will be applied to any classification model selected for use in this thesis, regardless of the level of the classifier itself. LIME is explained in further detail in the following.

3.1.1.1 Local Interpretable Model-Agnostic Explanations

In the paper presenting LIME, an explanation is defined as "presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction". The model used to provide the explanation is an interpretable model $g \in G$ where G is a class of model such as linear models, decision trees, or falling rule lists. Furthermore, as even these, otherwise interpretable models can become uninterpretable if a large number of features are combined in the model to an extent where it is no longer possible for a human to readily comprehend, a complexity term $\Omega(g)$ is used for the explainer model. For decision trees this is the depth of the tree and for linear functions it is the number of non-zero weights, thereby ensuring a more simple and interpretable model [35].

The explainer model works by analysing the information around the input of a specific prediction in a proximity denoted π_x. This is done by varying the input values within the proximity of x. The model for which the explanation is to be provided is denoted f, and the expainer model g then evaluates the output of the model f at the different inputs within the proximity. By doing so, the explainer is trained and can provide an explanation as to which input features are most important in providing the specific prediction[35].

The goal is then to optimise the explanation for interpretability and local fidelity. Local fidelity is a measure of how good the explanation is in the local area around the specific prediction and a measure for this is $\mathcal{L}(f, g, \pi_x)$. With the complexity term $\Omega(g)$ to ensure interpretability and the local fidelity measure, the explaination is provided by LIME as:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L} + \Omega(g). \tag{3.1}$$

A drawback to LIME is that when the explanation is based on a linear function it will not be able to provide an accurate explanation for a prediction if the local proximity is highly non-linear, and for some representations it may also not be possible to provide an explanation. In the paper presenting LIME [35], they use the example of a model predicting sepia-images as *retro*, this will not be possible to explain from the presence or absence of specific pixels in the image.

The benefit of the LIME approach for an explanation is that it is model-agnostic and does not make assumptions about the model to be explained.

3.2 Classification

3.2.1 Selection of classification algorithms

Selecting a classification algorithm is central to the development of a predictive model. There are different algorithms that can be applied to a binary prediction problem. In the present thesis the selection of the algorithms is based on two main criteria listed below:

- Develop a model that can provide transparency
- Select an algorithm with a possibility of achieving good predictive performance for the defined problems

Looking at studies for other clinical decision support systems, some machine learning models have yielded good results. Both Horng, Sontag, Halpern, *et al.* [11] and Lin, Chen, Tseng, *et al.* [13] have applied logistic regression and the more sophisticated random forest model with good results for both. In the study by Mao, Chang, Tsai, *et al.* [18] logistic regression was also applied. In Canada, K-nearest neighbors (KNN) was applied in the paper [20], and support vector machines (SVM) was applied and compared to KNN in [19].

Logistic regression is of interest as it is a relatively simple model, where the function coefficients can provide information about how decisions are made. This fits well with the goal of transparency. However, since the model produces a linear decision boundary, it is not possible to model complex relationships between predictors. Consequently, if the true underlying relationship is non-linear, a logistic model may not produce good results. However in many cases a linear representation may be sufficient. The results achieved by Horng et al. in [11] were comparable from a logistic regression model and the non-linear random forest model when predicting occurrence of sepsis in patient. In the study [13] of post-stroke activities of daily living, Lin et al. achieved good results with logistic regression, although a small, but significant, AUC difference in favor of a random forest model was seen. Furthermore, a systematic review [38] from 2019 showed no improvement in performance when more advanced machine learning models were applied to clinical risk prediction as opposed to applying logistic regression. To achieve transparency and lower computational complexity it is preferable to adopt a simple model when possible, it is therefore relevant to investigate the performance of a logistic regression model in the present thesis.

K-nearest neighbors require no training, but demands a rather large set of training observations for comparison at the time of prediction [39, pp. 463-468]. This means that an up to date set of data must be available for the model in production. For now all utilised data is anonymised, in which case this should not be a problem. However, if the developed model from this thesis in the future is to be updated with more data, it may be of interest to gain additional insights from data that has not been or cannot be anonymised in order to improve model performance, and in that case a KNN-model would not support this. This is considered a drawback of the method.

Random forest [40] and support vector machines are both more complex and less transparent methods than logistic regression. However, they have also yielded good results [11, 13, 19]. Furthermore, random forest is capable of defining a non-linear decision boundary, which is interesting if the underlying relationship between predictors is non-linear.

It is of interest to select both a method with a linear decision boundary and a method with a non-linear decision boundary as the underlying relationship of the problems investigated in the present thesis is not known. Furthermore, as logistic regression

is considered relatively transparent and has provided good predictive performance in comparable studies, this is selected as a model to be applied in this thesis. Additionally, random forest is also selected, as this is capable of representing a non-linear decision boundary and it has shown good results in other studies.

The selected methods are described in further detail in the following sections.

3.2.2 Logistic Regression

Logistic Regression aims to model the posterior probabilities of the classes via linear functions in x while still ensuring that they sum to one and remain in $[0, 1]$. These output probabilities can then be used to classify observations. In figure 3.2 an example of a logistic function with a one predictor as input, can be seen. This illustrates how the output probability is constrained to values between zero and one.

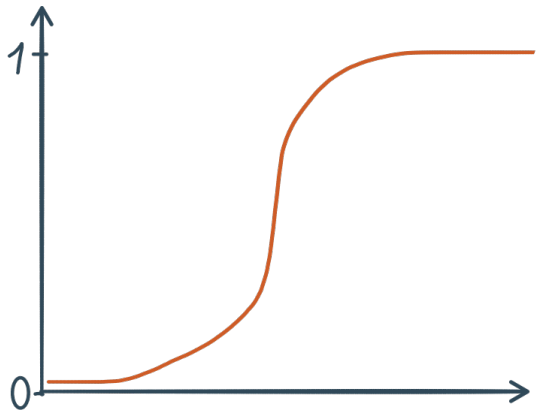


Figure 3.2: *An illustration of a logistic function for input in* $\mathbb{R}^1$.

For a multiple logistic regression applied to a binary classification problem the logistic function is a follows:

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}, \tag{3.2}$$

where $X = (X_1, ..., X_p)$ are the p predictors and $\beta_0, \beta_1, ..., \beta_p$ are the corresponding coefficients [39, p. 119] [41, pp. 135-136].

Logistic regression allows for some insights into the association between the predictors and the outcome. From the coefficients it is possible to see whether there is a positive or negative relationship between a predictor and the output. This comes from the property that the log odds of the probability are linear in x as seen in the following rewrite of the equation:

$$\log_e \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p. \tag{3.3}$$

This also means that logistic regression produces a linear decision boundary, which can be set at 0.5 probability or defined at a specific sensitivity or specificity in a particular application. As mentioned above, with a linear decision boundary, it is not possible to model complex relationships between predictors.

3.2.2.1 Fitting the model

The coefficients of the logistic model are estimated using the training data and maximum likelihood. The goal is to obtain coefficients values for which the estimated probabilities of each sample is as close to the true, observed class as possible. More specifically, it is the log likelihood which is maximised to fit the coefficients[39]. The log likelihood function for a two class classification problem looks as follows:

$$\ell(\beta) = \sum_{i=1}^{N} \Big\{ y_i \log p(x_i;\beta) + (1-y_i) \log 1 - p(x_i;\beta) \Big\}. \tag{3.4}$$

The log likelihood function is differentiable, thus, to maximise the function the derivatives are set to zero:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = 0. \tag{3.5}$$

The equation does not yield a closed form solution and is therefore solved iteratively. There different methods that can be applied to solve this. One category of optimisers are first order methods such as the method of steepest ascent, which use the first order derivatives. However, this can be slow to converge because of updates to the search direction at each step [42]. A faster converging alternative is Newton-methods which use the Hessian matrix, and makes updates as:

$$\beta^{new} = \beta^{old} - H^{-1}(\beta^{old}) G(\beta^{old}), \tag{3.6}$$

where $H^{-1}(\beta^{old})$ is the inverse Hessian of β^{old} and $G(\beta^{old})$ is the gradient of β^{old} [39]. However, this can be rather computationally expensive, which limits the use in practise. Often used methods are therefore the quasi-Newton methods, where the Hessian is approximated [42]. For this, the Hessian $H^{-1}(\beta^{old})$, is replaced with a local approximation of the inverse Hessian, B:

$$\beta^{new} = \beta^{old} - B G(\beta^{old}), \tag{3.7}$$

where B a symmetric, positive definite matrix.

One such method which has been further optimised to use less memory, is the L-BFGS-B algorithm [43]. Like other quasi-Newton L-BFGS-B uses an estimate of the inverse Hessian to guide the search of the maximum. The algorithm is not guaranteed to converge, but it has proven useful in practice, where it converges in relatively few iterations compared to other optimisers and with similar accuracy[42, 44]. L-BFGS-B algorithm is selected for use when fitting logistic regression in the present thesis.

3.2.2.2 Regularisation

As previously mentioned a challenge when developing a model is to achieve a high predictive performance while at the same time not overfitting to the training data, as the model should generalise to new data. When fitting a logistic regression model, there is a risk of overfitting to specific predictors, which the model may assign large coefficient values for. To avoid this, regularisation can be employed. This entails assigning a penalty to large coefficient values when conducting maximum log likelihood estimation. This can also help reduce unwanted skew in coefficients of predictors that have some correlation between them. The penalty term; $\lambda R(\beta)$, has a tuning parameter λ, which determines the strength of the regularisation, and uses a regularisation function R. [39, pp. 61-63, 662].

There are different types of regularisation that can be applied. Some common methods are L1-regularisation $||\beta||_1$ and L2-regularisation $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$, where β_0 for the intercept is not penalized. L1-regularization is capable of shrinking coefficients, but also removing predictors altogether by setting coefficients to zero, while the L2-regularization generally shrinks coefficients and rarely removes predictors completely[39, pp. 61-69, 662].

As subset selection is applied separately it is not of specific interest to remove predictors by regularisation, and furthermore, since the Euclidian distance, which the L2-regularisation is based, is differentiable, L2-regularisation may be a better choice as the optimizer selected uses an estimation of the derivatives to evaluate the result. Thus, the L2-regularization is selected for use, when estimating the coefficients of the logistic regression model.

3.2.2.3 Feature scaling

When applying regularisation, the features should be scaled so that the different predictors are on the same scale, since the coefficients are added together in the penalty term. This is done as standardisation by subtracting the mean and dividing by the standard deviation within each predictor:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}. \tag{3.8}$$

3.2.3 Random Forest

Random forest is a decision tree method for classification. Decision trees have multiple advantages which makes them fit for the classification task relevant for this thesis. They are simple to understand and can be visualised by white box modelling, meaning a full explanation can be obtained for for a given condition. However, their prediction accuracy is rarely competitive with other more advanced supervised methods [41, p. 303]. Additionally, decision trees are not robust. Often, a small change in the data impacts the tree significantly leading to a different series of splits, thus they are said to

have high variance. This is due to the hierarchical nature of the tree where the effect of a top split is propagated down to all splits underneath. These disadvantages aside, they form a foundation for random forests, which is a better performing classification algorithm [p.303][41] that applies various methods for improving the decision tree. An example of the structure of a decision tree is shown in figure 3.3. The decision tree

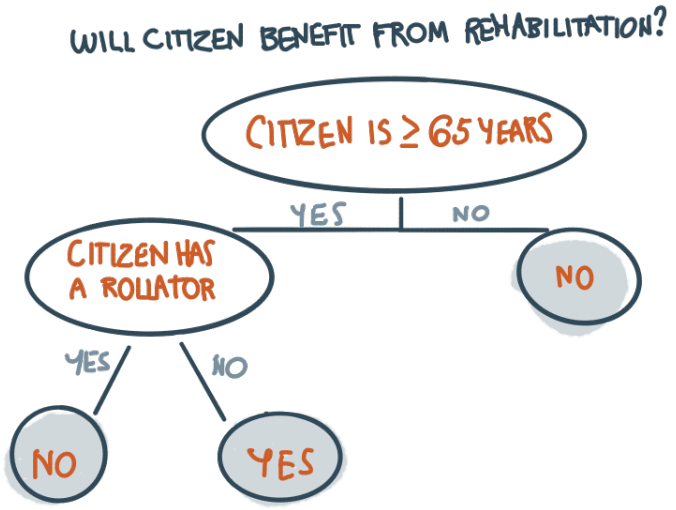


Figure 3.3: *A hypothetical example of a decision tree predicting whether a citizen will benefit from rehabilitation based on two predictors. Leaf nodes at the bottom of the tree are the resulting classification of the observation.*

for classification aims to predict a qualitative response. The observations are denoted $\{(x_i, y_i\}$ for each $i = 1, 2, \cdots, N$, where y_i is the true class label while x_i is a set of M features f, on the form $(x_{i1}, x_{i2}, \cdots, x_{iM})$. The algorithm partitions all observations greedily, starting with all observations a splitting variable j and split point s, a pair of half-planes are defined:

$$R_1(j,s) = \{X \mid X_j = llama\} \text{ and } R_2(j,s) = \{X \mid X_j \neq llama\} \tag{3.9}$$

The X_j and s that minimizes the criterion is selected using the Gini index or cross-entropy. The Gini index is a measure of the total variance across the K classes and is defined by equation 3.10 [41]. A low Gini index is obtained if all of the proportions of training observations in the mth region that are from the kth class $\hat{p}_{mk}$ are close to zero or one.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.10}$$

Cross-entropy defined in equation 3.11 is similar to the Gini index as a small value is obtained likewise.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{3.11}$$

The motivation for using the Gini index or cross-entropy compared to the classification error rate is found in them being more sensitive to changes in node probabilities. In a binary classification setting with 500 observations in each class denoted by $(400, 400)$, one split might create two nodes $(300, 100)$ and $(100, 300)$ while another split created nodes $(200, 400)$ and $(200, 0)$. Looking at the misclassification rate, this yields 0.25 for both splits. However, the Gini index and cross-entropy prefer the latter as a pure node is obtained. The selected measure for this thesis is the Gini index as the computation of the logarithmic function is avoided. In the first split of a decision tree all observations are parted into two regions R_1 and R_2. This is the optimal split and the splitting is continued recursively for each region. The tree now consists of a number of nodes each containing a decision rule that assigns observations to the child node. At each leaf node the observations are labelled to the majority class.

The process of growing classification trees can be regulated using hyperparameters. One of these is the depth of the tree, which also states the complexity. A shallow tree might not perform sufficiently in finding the patterns of the data while a tree that is too deep might fit to the noise of the data instead of the pattern. One approach to solving this is by defining a threshold, but this might not perform in cases where a seemingly worthless split is followed by an important split. Instead this is solved by growing a large tree T_0 and thereafter pruning the it to obtain a subtree T. N_m is the number of observations in node R_m given by $N_m = \#\{x_i \in R_m\}$ As there might exist a large amount of possible subtrees, cost complexity pruning is used, where a sequence of trees indexed by a tuning parameter α. For each value of α there exists a subtree T that minimizes the criterion. $Q_m(T)$ is the node impurity measurement - e.g. the Gini index as described previously.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T| \tag{3.12}$$

The random forest model was proposed by Breiman et al in 2001 [40] and is an algorithm using trees as building blocks to construct a more powerful prediction model. Random forests build a number of decision trees on bootstrapped training samples. This is also called bagging. By averaging these trees, this approach reduces the variance. Additionally, random forests decorrelates the decision trees by choosing a random sample of m predictors as split candidates from the full set of p predictors. This approach improves the results obtained by bagging, which might have a tendency to grow similar trees in the case of a strong predictor. Hence bagged trees might be highly correlated and thereby an average of these trees cannot reduce the high variance of the decision tree. If the strong predictor cannot be chosen in all of the top split of the subtrees, other predictors have more of a chance and this decorrelates the trees, making the average of the trees less variable and therefore more reliable. In the case where $m = p$, random forests and the bagging approach is similar. Using smaller values of m when building random forests will typically be helpful if many of the predictors are correlated. Typically, the number of features for each tree is calculated by the square root of

the total amount of predictors $m \approx \sqrt{p}$. The algorithm for random forest is outlined below. This grows B trees.

1. $b = 1$ to B:
   (a) Draw a bootstrap sample **Z*** of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminated node of the tree, until the minimum node of size n_min is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$. To make a prediction at a new point x:

Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote \left\{\hat{C}_b(x)\right\}_1^B$.

The prediction is based on majority vote among the trees. The implementation of random forest used is part of the Scikit-learn library [45].

With regards to the transparency of the model, it is possible to derive a set of feature importances from the trained model. The approach used by the Scikit-learn implementation to calculate the feature importance, is by observing how random re-shuffling of each predictor affects the model performance [46]. If the model has suffered from overfitting to the training data, this means that the importances might be high for features that are not predictive of the target value. The optimum solution to this would be to derive the feature importances from the hold-out set.

3.2.4 Hyper parameters

A commonly used method to improve the performance of classifiers is to apply hyper parameter tuning. Different properties of the classification methods can be tuned to achieve models better fit for the prediction of certain data. Two methods that can be used are *grid search* and *random search* [47]. Both methods are based on combining different values for the different hyper parameters and evaluating model performance to select the optimal values. Grid search uses a structured approach to in selecting values, whereas random search, as the name indicates, uses a random approach. Random search has been proven superior to grid search in optimising the tuning parameters [47].

Although a great potential for improvement in the predictive performance may be achieved from tuning and optimising hyper parameters, this is left to future work in the present thesis, as the focus regarding optimisation of the predictive performance is pertained to exploring the data in order to create and select useful predictors and determine how information about assistive technology can be leveraged in the models.

3.2.5 Feature selection

Feature selection is used to find the combination of parameters that optimises the model's prediction capability. It is highly important that noisy and redundant features are removed. This will provide a higher performance for the classifiers and eventually lower the feature dimension leading to a more lucid prediction. Various methods can be applied to find the most important subset of features. The best subset selection is one method that is designed to fit all possible combinations of features for a model [41, p. 205]. However, predicting with all combinations of a feature set is a resource intensive challenge. The more lightweight approach to this is by a stepwise selection of features, where features are either added to or removed from a set of features, one at a time. In forward stepwise selection the model searches through the features to find the one yielding the best score. This continues greedily until the score does not improve. In backwards stepwise selection the model starts with all features and greedily removes one at a time to find the optimum [39, p. 58]. Backwise stepwise subset selection can only be used when the number of observations is larger than the number of features. Another approach is by using a tree algorithm, which has built-in measures for feature importance applicable for selecting the feature subset. Decision trees use the Gini index for this calculation as described in section 3.2.3. In this thesis, forwards stepwise subset selection is used as it is widely applicable and provides a fair estimate of the best features [48]. Furthermore it is constrained and does not require the computational costs as best subset selection.

3.3 Evaluation metrics for binary classification

When training a classifier there are basic outcome values from the predictions that are relevant for further evaluation metrics. For a binary classifier these can be presented using a confusion matrix as seen in figure 3.4 on the facing page.The confusion matrix of a binary classifier contains four values. These depend on the predicted and the true conditions of a classification. Along the diagonal from top left to bottom right, the correctly predicted observations are presented, these are the true positives (TP) and true negatives (TN) results. The two remaining values are the false positives (FP) and the false negatives (FN). These are the incorrectly predicted observations.

From the values of the confusion matrix different evaluation metrics can be calculated. Two often used metrics are accuracy and precision [49]. These are calculated as shown in equations 3.13 and 3.14.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.13}$$

and

$$precision = \frac{TP}{TP + FP} \tag{3.14}$$

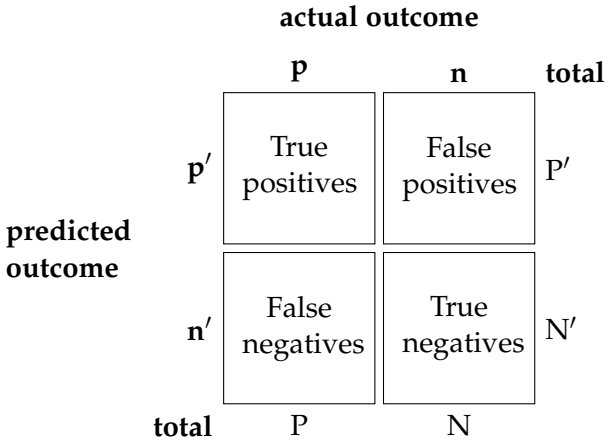


Figure 3.4: *Confusion matrix for evaluation of a binary classifier. Inspired from [49].*

Accuracy describes the fraction of correctly labelled observations from all observations, while the precision provides the fraction of true positive predictions from the total number of positive predictions and thus provides a measure of how useful the positive predictions are.

Additionally, two other interesting metrics that can be derived from the values of the confusion matrix are *sensitivity* and *specificity*. Equations 3.15 and 3.16 show the calculation of these.

$$sensitivity = \frac{TP}{TP + FN} \tag{3.15}$$

$$specificity = \frac{TN}{FP + TN} \tag{3.16}$$

These each depend on the values of one column in the confusion matrix. The sensitivity is also called recall or true positive rate. This defines the models probability of correctly labelling the positive observations as such, while the specificity is the probability of correctly labelling the negative observations. Related to the specificity is the false positive rate as seen in equation 3.17.

$$false\ positive\ rate\ (fpr) = 1 - specificity = \frac{FP}{FP + TN} \tag{3.17}$$

While accuracy and precision are often-used metrics, these are in many cases not optimal for performance evaluation of prediction models [49]. One problem arises because the metrics evaluate both positive and negative samples together. This makes them easily affected by skew. As an example a model trained on an imbalanced data set, can achieve a high accuracy by classifying all samples to the majority class, but this does not make the model useful. Additionally, variations in the class skew can also affect these metrics.

3.3.1 The receiver operating characteristic (ROC)

The receiver operating characteristic curve is a graph used to visually represent the performance of a binary classifier at different thresholds. In figure 3.5 an illustration with different examples of ROC curves can be seen. At the x-axis of the graph is the false positive rate which corresponds to $1 - specificity$, and at the y-axis of the graph the true positive rate (*sensitivity*). A ROC curve of a model starts in (0,0) which represents a threshold where no false positive results are obtained, but this also entails no true positive predictions as all predictions will be negative. At the other extreme at the point (1,1) only positive predictions are output [49]. Depending on the domain and application the threshold can be varied to achieve an acceptable trade-off between the two measures. An optimal classifier with perfect prediction would achieve a point at (0,1) with no false positives and 100% true positives, whereas a curve following the diagonal line from (0,0) to (1,1) does not perform better than random guessing [49].

Figure 3.5: *Illustration of the ROC space with examples of different ROC curves corresponding to different classification performances.*

For classifiers that only output class labels, only a single point on the graph can be calculated as the threshold cannot be varied according to probabilities of the predictions. However, for classifiers with the possibility of producing a predicted probability varying the threshold of either the false positive rate or the true positive rate will allow for multiple points on the graph to be calculated [49].

3.3.2 The area under the receiver operating curve (AUC)

To compare and evaluate models based on the ROC curve, it is useful to obtain a single value as the evaluation metric. This is achievable by calculating the area under the ROC curve (AUC). The AUC value has the interesting property, that for a random

pair of a positive and a negative sample the AUC represents the probability that the positive sample will be classified as positive with a higher certainty than the negative sample. [50]. There are further properties of the AUC that makes it highly suitable as a performance measure for prediction models. One is that it has been proven to be more sensitive than accuracy achieving higher levels of significance levels when compared. In addition to that, AUC is not affected by the decision threshold and independent of any prior probabilities in the classes [51].

Two different approaches to computing AUC can be used. Firstly it can be calculated from a ROC curve fitted using iterative Maximum Likelihood estimation where the slope and intercept are used to obtain an estimate of the AUC. This assumes a Gaussian distribution for the underlying probabilities of prediction. Secondly, it can be calculated using trapezoidal integration to estimate the AUC. The drawback of the trapezoidal integration is that it systematically underestimates the area. However, this error can be reduced by ensuring a larger number of points on the ROC graph [51]. Furthermore, as this underestimation is present across all models, the comparison of models will be less affected.

3.3.3 Threshold

While the ROC curve displays the FPR and TPR trade-off at different thresholds, it is common to select a specific threshold which reflects the priorities for the specific domain and application in which the model is to be utilised. This threshold can be in terms of either specificity or sensitivity and then the other measure can be optimized according to this.

3.3.4 Selection of evaluation metrics

The selected metrics must be appropriate for the specific application. In the present thesis, some models are trained on data with an uneven proportion of class labels. Thus, it is important to have an evaluation metric that is independent of class skew. For this reason accuracy and precision are not suited as metrics for the comparison across different models.

Furthermore, in the domain of the health care sector it is important to avoid wasting scarce resources providing rehabilitation programmes to citizens, who will not benefit from it. It is, however, also highly important to ensure that any citizen who will benefit from it is provided the opportunity to enroll in the program.

As metric for optimising the model performance AUC is selected since it is resistant to skew and optimises across all thresholds, which is acceptable. For calculation of the AUC value, the trapezoidal method is selected, since this approach does not require any assumptions about the underlying distributions.

3.4 Model selection and assessment

This section aims to describe the methods for performance assessment and how they are used to select models. Model selection and assessment are two methods used in order to firstly select the model and thereafter evaluate it. Thus it presents two distinct objectives [39, p. 222]:

- **Model selection:** Estimating the performance of different models in order to choose the best one.

- **Model assessment:** Having chosen a final model, estimating its prediction error generalisation performance) on new data.

In general, for a statistical learning model, a large amount of data is desired as it aids in reducing the signal-to-noise ratio. However, it is of no interest to train and evaluate the model on the same data set since it would typically introduce an optimistic bias due to overfitting and the model's prediction capability on an independent data set cannot be correctly estimated. This is referred to as the Bias-Variance trade-off and illustrated in figure 3.6.

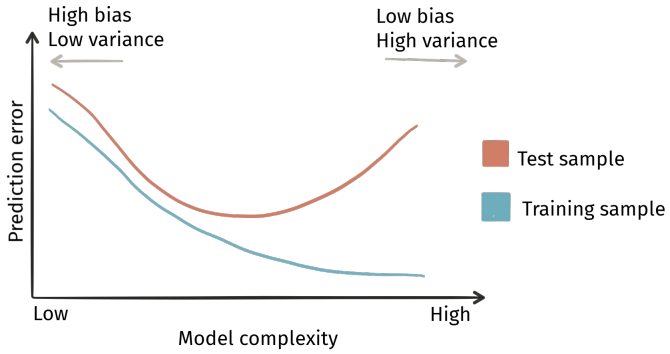


Figure 3.6: *Comparison of training and test error with increasing model complexity.*

The overall objective is to perform model selection and model assessment while ensuring that the results for these are unbiased. Therefore, before applying any feature subset selection or classification algorithms, the data should be partitioned to ensure that model selection and assessment are kept separate [39, p. 222]. Such a split is illustrated in figure 3.7.

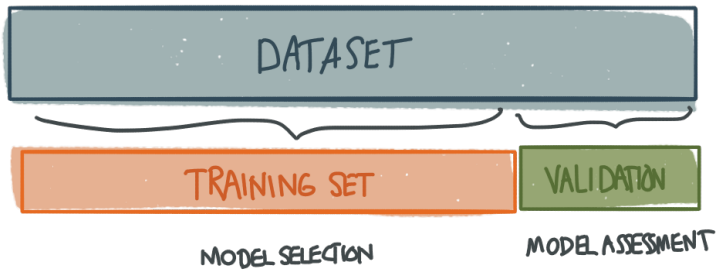


Figure 3.7: *Validation set approach. The data set is partitioned into two parts of varying size.*

A typical size of the validation set is $20 - 25\%$ depending on the size of the data set [39, p. 222]. By creating the validation set, the unbiased performance of the selected best model can be evaluated. The drawback of this approach is that it lessens the amount of data which could have been used for optimizing the model during model selection, thus the model might perform worse. The next sections will describe the process of model selection and model assessment respectively.

3.4.1 Model selection

The training data is used for selecting the best classifier. Being in a data-rich situation, one could perform a partition of the training data to create a second training and testing set. This is the most computationally effective method. However, removing even more data from the training set could be problematic in situations with smaller data sets. The most widely applied method to approach this situation is by *k*-fold Cross-Validation [39, p. 241]. Cross-Validation is a resampling method used to provide an estimate for the prediction error. It works by splitting the training set into *k* partitions of approximately equal size. The first fold is treated as a validation set and the algorithm is fit on the remaining *k-1* folds. This is repeated *k* times, each time shifting the validation fold to another partition as shown in figure 3.8.

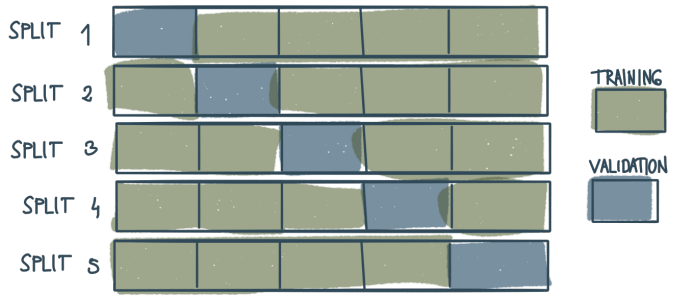


Figure 3.8: *Illustration of a 5-fold Cross-Validation.*

The motivation for this is again the bias-variance trade-off, illustrated in figure 3.9. Generally, a 5-fold or a 10-fold performs well [41, p. 183].

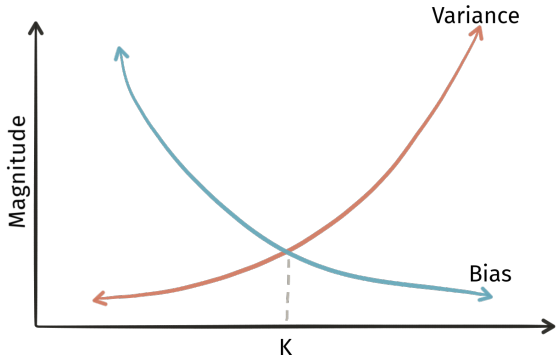


Figure 3.9: *Bias-Variance trade-off associated with k*

Combining Cross-Validation with AUC scores can be approached in two incompatible ways [52]. One of these is to sort the individual scores from all folds together to plot a single ROC curve and then compute the area under this curve. By sorting different folds together, it is assumed that the classifier should produce well-calibrated probability estimates. If calibration or specific threshold values are of no interest to the overall model selection, the classifiers will be unnecessarily downgraded. The AUC under Cross-Validation in this thesis will be calculated by computing the AUC for each fold and then average over the folds as shown in equation 3.18.

$$AUC_{avg} = \frac{1}{k} \cdot \sum_{i=1}^{k} AUC^{(i)} \tag{3.18}$$

If a fold would contain no positive labels, the AUC cannot be computed. To avoid this situation, a special kind of Cross-Validation is used in this thesis. This is called stratified Cross-Validation and ensures an equal distribution of class labels within each fold.

3.4.2 Model assessment

The goal of model assessment is to obtain the models capability in predicting on a independent test set - also referred to as the generalisation performance. Therefore the selected model is used to predict on the validation set as illustrated in figure 3.7. The AUC and ROC curve can be obtained from this prediction to estimate the generalisation error for the model and to finally conclude on how the model performs.

4

Data analysis

This chapter describes the data received from the industrial partners, Aalborg Municipality and DigiRehab. The aim is to gather an understanding of what the data contains, how the data is collected and how it is utilised throughout this master's thesis. This is partly done in collaboration with DigiRehab, as understanding the health and home care domain is vital for a correct utilisation of the data. Furthermore, this chapter describes how the data is prepared to fit the thesis objectives. This includes the data preparation steps such as data cleaning and filtering. Data preprocessing is important because the data will have to be transformed to optimise it for the applied statistical and machine learning methods [53, 54]. This entails careful handling of irregularities and missing values in the data.

4.1 Data sources

The collected data is retrieved from two independent databases that each contain information about citizens in Aalborg Municipality. These are KMD Nexus and DigiRehab. An overview of the data sources and the data provided can be seen in figure 4.1.

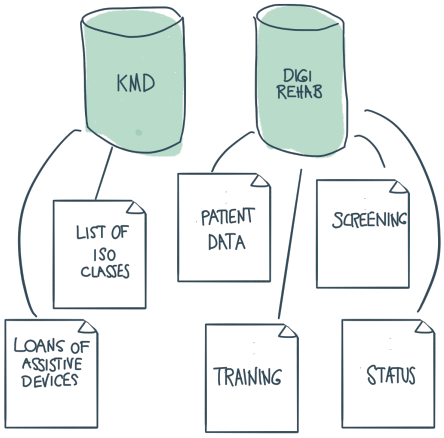


Figure 4.1: *The structure of the original files from KMD and DigiRehab.*

KMD provides data about loans of assistive devices. These files are described in section 4.2.2 DigiRehab has made their data about physical rehabilitation programmes of citizens available. This includes training data, screening information, patient data and status data. These will be described in further detail in the following sections. The available data is structured in a set of comma separated files. The two data sources share a subset of citizens which is used in this thesis as illustrated in figure 4.2.

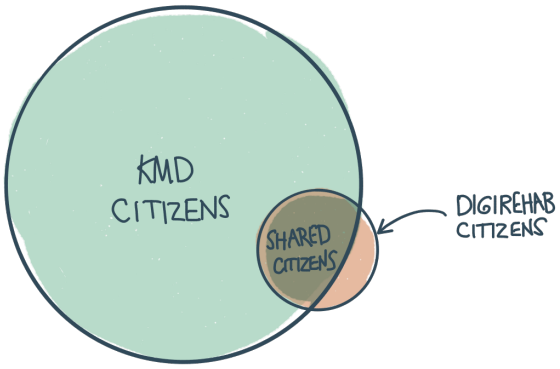


Figure 4.2: *Illustration of the citizens within each data set.*

4.2 Data structure

In this section the structure and content of the data is investigated in order to understand the possibilities and limitations of the information. Each section below describes the data source and the different files as shown in figure 4.1.

4.2.1 DigiRehab application

DigiRehab has developed a mobile application used by health care workers to exercise with citizens receiving home care. Health care workers input the progress of citizens into the application, which generates tailor-made exercise programmes for each citizen based on their data. This implies an extensive data gathering which is used to evaluate on each citizen's progress simultaneously as they follow the programme. This means that the training programmes are continually adapted to each individual. Furthermore, from the data recorded the municipalities can obtain insights to the effect and impact for both the individual citizen and on a group level for all citizens enrolled in the system. DigiRehab is currently in use by 19 of the 97 Danish municipalities. One of these is Aalborg Municipality. At the initiation of the collaboration between Aalborg and DigiRehab a test period and preliminary evaluation was conducted to assess the value and effect of DigiRehab rehabilitation programme [55]. For the assessment 75 citizens were enrolled in the program. 47 citizens completed the programme and their average time of home help was reduced by 88 minutes per week. The reduction of time was still apparent 12 months later. In August 2019, DigiRehab completed an impact analysis in

Ballerup Municipality for 57 citizens during a nine months interval between August 2018 and May 2019 [23]. This analysis describes how DigiRehab utilises the various information from the data to assess the citizen's potential for improving during a rehabilitation programme. Out of the 57 citizens, 45 completed the programme by training in 8 of the 12 weeks. DigiRehab evaluates this as on a par with the national average. The succeeding sections describe the data received from the DigiRehab application.

4.2.1.1 Patient Data

This file provides an overview of the citizens who have been enrolled in a DigiRehab rehabilitation programme. DigiRehab refers to citizens as *patients*. An overview of the contents of the file is given in table 4.1.

Entry	Description
Patient ID	A unique identifier for the patient.
Citizen ID	The corresponding ID for the citizen used in the KMD database.
Gender	The gender of the patient.
Birth year	The birth year of the patient.

Table 4.1: *The patient data file.*

This file ensures that citizens in DigiRehab data can be found in the KMD data, as it provides the identifiers used in both databases. The file consists of 649 unique citizens of which 413 are women and 211 are men. 25 of the entries are missing the gender and birth year. They are in average born in 1935.

4.2.1.2 Screening

This file contains the information gathered when DigiRehab conducts a screening of a citizen. The purpose of a screening is to map the citizen's functional capacity and create an exercise programme tailored to the citizen. As additional screenings are made, it is possible to follow the progress in terms of the citizen's self-dependence and physical capacity. The first half of the screening consists of ten questions regarding the citizen's need for help in their daily life. The second half includes ten physical tests of the citizen's strength. The outcome of a screening is illustrated in figure 4.3.

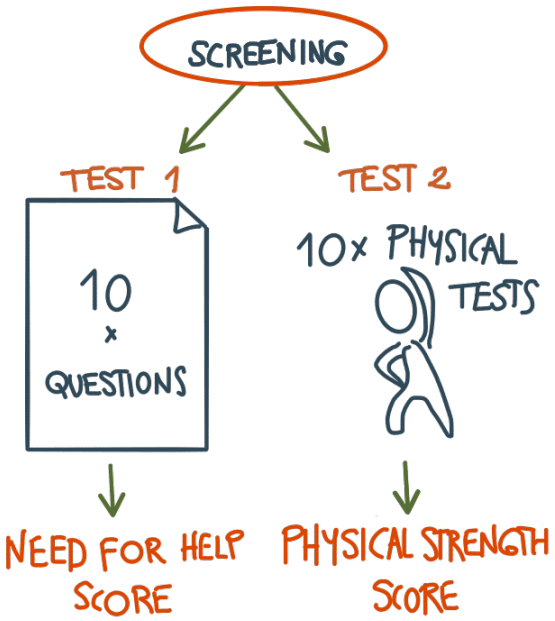


Figure 4.3: *Illustration of the screening procedure.*

The contents of the file is described in table 4.2

Entry	Description
Patient ID	A unique identifier for the patient.
Screening date	The date for the screening.
Need for help score	A measure of the citizen's need for help in daily activities. This value is between zero and 100.
Physical strength score	A measure of the citizen's physical strength. Based on the ten physical tests conducted at the screening. This value is between zero and 100.
Exercises	A list of exercises customised for the citizen.

Table 4.2: *The screening file.*

On basis of the screening, the citizen receives two scores between 0 and 100, one outlining their *need for help* and one indicating their physical strength. If the citizen is self-dependent, the *need for help* score is low and vice versa. Physical decline is described by a low *physical strength* score. The screening also yields a set of exercises which is customised to the citizen based on the test results. These are given as a list of numbers in the data. If a citizen is screened and one of the two scores has worsened since the last screening, the health care worker is prompted to fill out a reason among a set of the values listed below.

- Generelt ustabil *(Generally unstable)*
- Sygdom *(Sickness)*
- Fald/uheld *(Fall/accident)*
- Ingen forklaring *(No explanation)*
- Andet *(Other)*

The screenings are conducted before a rehabilitation programme is initiated and then again at regular intervals during the programme; optimally every four weeks. However, for the received data, the time between screenings ranges from 0.1 to 125.9 weeks, with a mean interval between screenings of 8.2 weeks.

4.2.1.3 Training

This file contains information regarding the citizen's trainings. The contents of the file is described in table 4.3.

Entry	Description
Patient ID	A unique identifier for the patient.
Training date	The date for the training.
Rating	The rating score (between 0 and 6).

Table 4.3: *The training file.*

As defined by DigiRehab [23], the citizen should complete at least eight trainings over a 12 week period for the programme to be optimum. The citizen is guided by the health care worker throughout the exercises and each training session is concluded by a rating score from 1 to 6, where 1 implies that the training was cancelled. A rating between 2 and 6 describes to which extent training with the citizen is considered meaningful. So far, this is the first measure which is purely subjective and based on the helper's experience with the citizen. Different helpers might have distinct understandings of how to use this score. If the helper wants to, comments can be added explaining their training session or reason for the rating. However, to ensure full privacy, the comments are removed from the data in this thesis as they might include names or other identifiable information. In total, 854 trainings were cancelled while 10,331 were completed. The mean rating for the completed trainings was 3.95.

4.2.1.4 Status

At all times the citizen has a status which explains their current course in the exercise programme. This status is chosen by the health care worker manually from a set of predefined values. This set consists of the following values:

- Aktiv *(Active)*
- Pause *(Pause)*
- Afsluttet *(Terminated)*
- Afdød*(Has died)*
- OBS *(Monitor)*
- Vedligehold *(Maintenance)*
- Genaktiveret *(Re-activated)*
- Fravalgt *(Opted-out)*
- Markant fremgang *(Significant progress)*
- Venter *(Waiting)*
- Skal screenes *(Should be screened)*
- Auto-afsluttet *(Automatically terminated)*

The contents of the status file is described in table 4.4.

Entry	Description
Patient ID	A unique identifier for the patient.
Date	The date for the status change.
Status	The status for the patient.

Table 4.4: *The status file.*

The status file consists of 1,496 entries for 613 unique patient IDs. The most commonly used status for the received data is *Terminated* which has been used 570 times. 100 citizens have been given the *Has Died* status.

4.2.2 KMD Nexus

KMD Nexus is a digital healthcare platform used by municipalities to record and evaluate on citizens receiving home care. The application supports stock control of all assistive devices owned by the municipality which is out on loan - both previously and currently.

4.2.2.1 Loans of assistive devices in Aalborg Municipality

This table consists of data on the assistive aids currently and previously lent by citizens in Aalborg Municipality. Each device has an HMI-number and an HMI-name which corresponds to a product ID and name. Devices hold an article number which is a unique identifier, a paragraph and dates for the loan and the possible return of the device. In total the file contains 609,948 entries of loans related to 56,674 different citizens. The contents of the assistive devices file is described in table 4.5 on the facing page.

Entry	Description
Citizen ID	A unique identifier for the citizen.
HMI number	The product number.
HMI name	The product name.
Device ISO class	The ISO class number of the device.
Article number	The serial number of the device.
Paragraph	The paragraph under which the assistive device was lent.
Date of loan	The start date for the loan.
Date of return	The end date for the loan.
Price	The price for the device.

Table 4.5: *The loans of assistive devices file.*

4.2.2.2 List of ISO classes

Along with the file for loans of assistive devices, a file describing the ISO classes was provided. The assistive products are classified in an order based on the international standard for classification of assistive products, ISO 9999. A separate file contains information about the currently valid assistive technology ISO-classes. The ISO-class numbers are structured hierarchically at lengths up to eight digits (four levels) where the grouping is more detailed the longer the ISO-class number is. This makes it possible to vary the level of detail used in processing. An example to illustrate this is provided in figure 4.4.

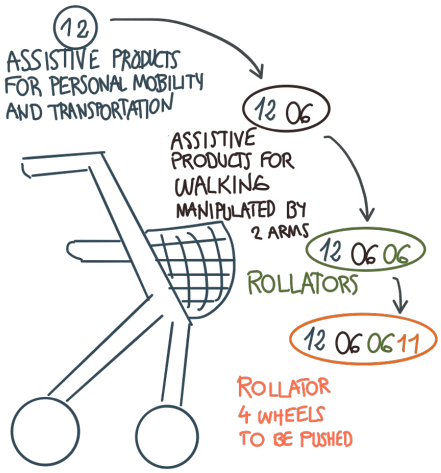


Figure 4.4: *Illustration of hierarchical structure of the ISO class for a 4-wheeled rollator.*

According to AssistData [56] there exists a total of 12 distinct categories each defined by two digits. The full list of categories is displayed in table 4.6 on the following page.

Category	Description
04	Assistive products for measuring, supporting, training or replacing body functions
05	Assistive products for training in skills
06	Orthoses and prostheses
09	Assistive products for self-care activities and participation in self-care
12	Assistive products for personal mobility and transportation
15	Assistive products for domestic activities and participation in domestic life
18	Furnishings, fixtures and other assistive products for supporting activities in indoor and outdoor human-made environments
22	Assistive products for communication and information management
24	Assistive products for controlling, carrying, moving and handling objects and devices
27	Assistive products for controlling, adapting or measuring elements of physical environments
28	Assistive products for work activities and participation in employment
30	Assistive products for recreation and leisure

Table 4.6: *The full list of categories for assistive products.*

The insights concerning the interpretation of the data gathered in this chapter serve as the baseline for the subsequent data processing described in detail in the succeeding section.

4.3 Data filtering

This section elaborates on the irregularities found in the data sets provided by DigiRehab and KMD and how these are addressed. Outliers and invalid data are carefully described in order to discard what seems to be faulty data entries. It is important to understand the impact of this, and to ensure the choices made in the preprocessing supports the defined problems of the thesis. To support the understanding, a diagram showing all of the processing and preparation steps can be seen in figure 4.5 on the next page. Descriptions concerning each step are found in the subsequent sections.

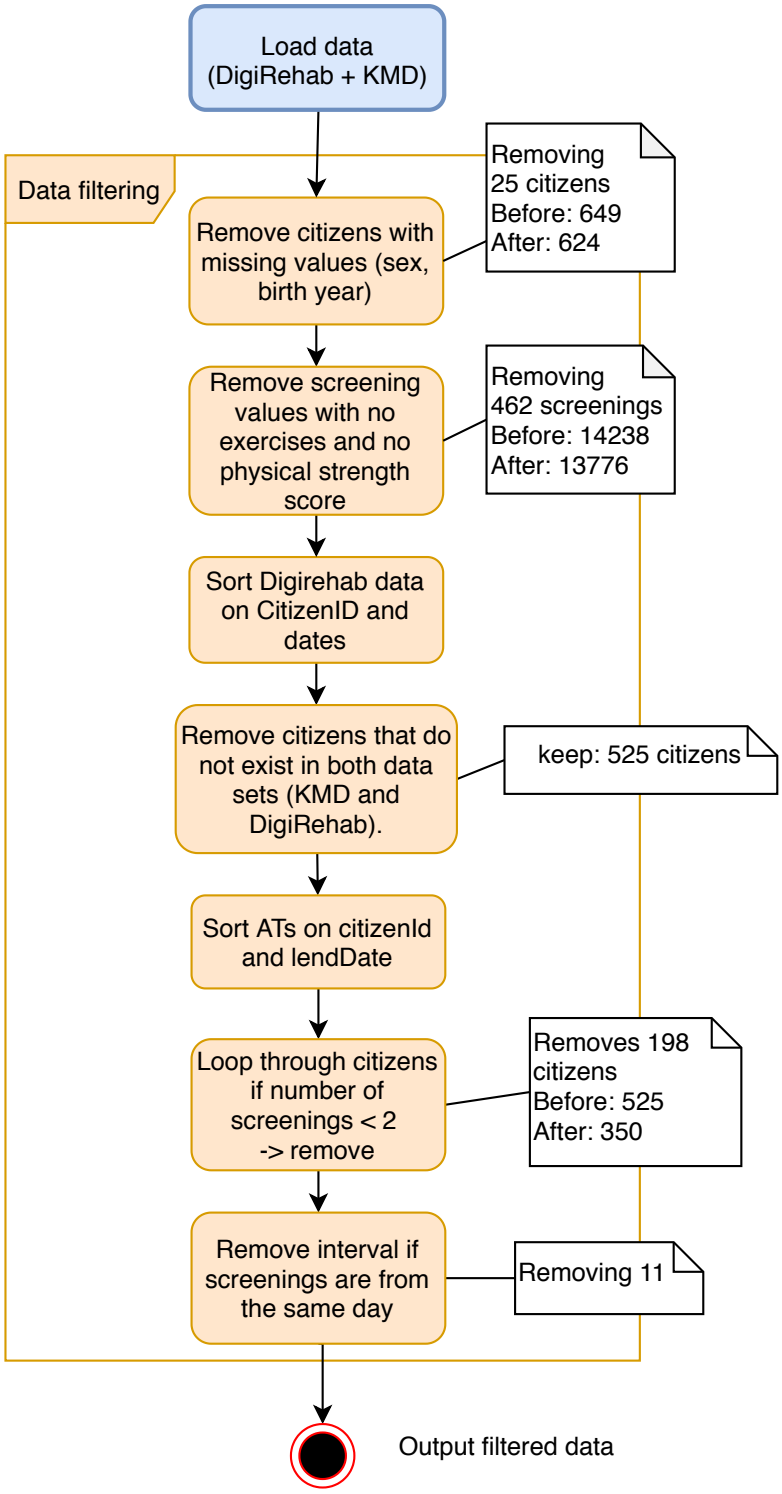


Figure 4.5: *The data is preprocessed and prepared as shown before being any models are applied.*

4.3.1 Handling missing values in the data

Missing values can be handled in different ways, usually by imputation [11, 16, 57] or by omission. In the data sets from DigiRehab, there are missing values in some data entries. The following sections outlines how these are handled.

4.3.1.1 Basic patient information missing

For the data provided by DigiRehab, there are 24 out of the 648 citizens in the Patient data file where sex and birth year is not provided. Additionally, the KMD Citizen IDs of these citizens is zero. Because the missing data in this case is considered important basic information for identification of the citizen, the omission approach is chosen, and these citizens are removed from the data. $(n = 624)$.

4.3.1.2 Screening entries with no exercises

Another file containing data irregularities is the screening values file. Each row in this data set corresponds to a screening where a need for help score and a physical strength score are provided along with a set of exercises for the citizen to do during the next training interval. The data contains 462 rows where no exercises have been assigned and furthermore, for most of these rows the physical strength score is also missing. This project is focused on evaluating who will benefit from rehabilitation, and since the rehabilitation is based on exercises assigned at screenings, it is not relevant for this study to look at screenings where no exercises are assigned. It is thus decided that any rows of screening data where the exercises field and/or the physical strength score is empty are removed from any further processing. The removed rows make up 3.2% of the 14,238 data entries in the screenings file.

4.3.2 Combining data from the two data sources

To use the data from both data sources it is necessary to match the data on IDs. In the patient data file, each citizen is assigned a DigiRehab patient ID and the corresponding KMD ID. The aim is to investigate who will benefit from rehabilitation and only the citizens present in the DigiRehab data have been part of the rehabilitation programme. For this reason, any entries in the loans of assistive techonology data is removed if the associated KMD ID does not belong to any of the 624 citizens in the DigiRehab patient data.

Furthermore, there are also citizens who are present in the DigiRehab data set, but not in the KMD data set. This might be explained by citizens that are newcomers to the municipality or simply have not loaned any assistive technology. Recall that the signature project aims to find correlations between rehabilitation programmes and loans of assistive technology as described in the Introduction, section 1.3. This thesis aims to provide a baseline for the subsequent work on the signature project which is why this work also has an interest in finding links between the two data sources. Therefore, if a citizen has no assistive devices and has never lent any, it is not possible to evaluate

how including or excluding the information about assistive technology will affect the predictive performance, and these subjects are therefore not included. Only citizens that are present in both data sets are included in the study. $(n = 525)$.

4.3.3 Handling citizens with no complete screening intervals

The citizen is continually screened during the programme. However, some citizens only have a single initial screening and no further screenings. As the evaluation of a screening interval depends on the development in the *need for help* score from one screening to the next, it is not possible to evaluate the effect of rehabilitation if a citizen does not have at least two completed screenings. It is therefore decided to remove citizens from the data if they have less than two screening entries present in the data, or if the citizen has two screenings recorded on the same day and no other screenings. This removes 175 citizens. $(n = 350)$.

4.4 Data preparation

To prepare the data set for model training and prediction, the representation and transformation can greatly impact the predictive performance of the model, and feature engineering is regarded as highly important to achieve good results for machine learning applications [53, 54]. Feature engineering includes creation and selection of features. This can be done manually by leveraging domain knowledge to create and select specific predictors, but it can also be done by creating a large number of candidate features and then using feature selection methods, the features appearing to be the most useful to the model are selected to be included in the final model [54]. As described in section 3.2.5 this thesis employs the latter approach.

4.4.1 Defining the target variables

The target variables used in the study have to support the definition of benefit defined in the Introduction, section 1.3.2. Since two separate objectives are investigated, two different versions of the response variable are created; one evaluating whether a citizen benefits from rehabilitation based on definition 1.1, where benefit is defined as the citizen having a decrease in the *need for help* score. The other is based on definition 1.2, where the aim is to predict who will complete a programme. The following sections outlines the choice of the response variables as derived from the data.

4.4.1.1 Predicting the development in the need for help score

Firstly, evaluating whether a citizen benefits from rehabilitation is in definition 1.1 described as the citizen having a decrease in the *need for help* score. As previously mentioned, this score ranges from 0 to 100 and is a measure of the self-sufficiency of a citizen. This measure is tightly linked to the amount of home care received by the citizen. The *need for help* score of a citizen is recorded at each screening. Evaluating the

development in this can therefore be done by comparing the *need for help* score from the start of the programme and again after a rehabilitation programme is complete after 12 weeks to see if the score has increased, decreased or is stagnant.

As the objective is to provide a classification method for predicting benefit vs. no benefit, the *need for help* score is transformed to a binary response variable, where a threshold is set and the discrete development of a citizen determines whether a positive label or negative label is assigned.

Setting the threshold requires making a decision about how much the *need for help* score must have decreased. As there is no control group, it is not possible to determine an improvement by comparison to this. Furthermore, as this study is completed in collaboration with external partners and the results should provide useful insights, an investigation of more than one threshold is of interest. This will also provide a glimpse of how a different threshold impacts the predictive performance. It has unfortunately not been defined by the collaborators how much improvement one should expect based on a rehabilitation programme. A consequence analysis conducted by DigiRehab [23] provides a small insight into the general improvement, but unfortunately the analysis is based on a slightly redefined *need for help* score. Therefore, the analysis does not provide an applicable benchmark. It is therefore determined to define three individual thresholds and provide results for each of these, such that the subsequent work can use this as reference points.

When setting the thresholds, it is important to remember that an increase of the score corresponds to the citizen being more reliant on the home carer than before the start of the rehabilitation programme. A decrease indicates the opposite. A stagnant development could be understood as the citizen not having benefit from rehabilitation. However, this conclusion might be naive as ageing usually leads to a continuous impairment of the physical condition as described by an effect analysis drafted by DigiRehab [23]. This is also supported by another study from 2013, which examines the hallmarks of ageing and defines it as a time-dependent progressive loss of the individual's physiological integrity, which eventually leads to deteriorated physical function [58]. It is therefore assumed that if the *need for help* score has not increased during the course of the rehabilitation programme, then the citizen must have gained benefit from the programme.

The first threshold is thus defined such that citizens who experience no change or have a decrease in the need for help score will be labeled positive and citizens that experience an increase in the the need for help score will be assigned a negative label. The class labels indicate that 53.6% have had benefit from the rehabilitation while 46.4% have not, so this yields a balanced data set.

For the second and third threshold the requirements for a positive label are slightly more strict. These are given at a decrease in the *need for help* score of at least 4 and at least 8. This means that a citizen must reduce their *need for help* score by at least 4 or 8 points to be assigned a positive label. These values are chosen with regards to the class distribution. An increase of 4 alters the distribution to 44.0% of the citizens having had benefit from the rehabilitation, while 56.0% have not. With a threshold of 8,

36.0% have had benefit whilst 64%. Choosing larger thresholds would lead to a steadily increasing majority class and result in an imbalanced data set. If we are to train a binary classification model without taking measurements for this problem, the model will be biased [59]. Therefore, due to the scope of this thesis, the thresholds are conservatively set to avoid the imbalanced learning problem.

4.4.1.2 Predicting who will complete a successful programme

When predicting who will complete a rehabilitation programme the response variable will also be binary to accommodate the objective. A successful programme is defined by DigiRehab as completing training sessions in at least eight out of 12 consecutive weeks. A citizen will therefore be assigned a positive label if this is achieved.

4.4.2 Tailoring the data to multiple objectives

Multiple experiments are conducted in this study to investigate the two objectives defined in the problem formulation, and these require different input. To accommodate this, three separate data sets are created from the filtered data. They are denoted feature vector 1, 2, and 3. The succeeding sections outlines the purpose and motivation of each feature vector and following this, an example with three fictional citizens is provided to increase understanding.

4.4.2.1 Feature vector to predict the development in the *need for help* score

Evaluating the change in the *need for help* score requires a screening 12 weeks after the first screening, as the rehabilitation programme has a total duration of 12 weeks. This means that for citizens that drop out of the programme early, the *need for help* score cannot be evaluated. To avoid excluding citizens that do not have a screening exactly 12 weeks after the programme start, any citizen with a screening within 12-26 weeks after en rolling in the programme will be included. 26 weeks is selected as the upper bound as observations from DigiRehab presented in [23], states that a positive development in the *need for help* score was still present after half a year (26 weeks). If a citizen has multiple screenings within this interval, the earliest screening is chosen. ($n = 125$).

4.4.2.2 Feature vectors for evaluating who completes a rehabilitation programme

For the second objective of predicting who will complete a successful programme, dropouts can be included and if they drop out early, they will simply be assigned a negative label. Thus, feature vector 2 includes any subject with two screenings. (n = 350).

It is also investigated how the predictions of a successful programme are affected when including information from the first few weeks of training. This is done as it is of interest to discontinue a rehabilitation programme if it is going nowhere. Feature vector 3 is created for this purpose.

DigiRehab aims for conducting screenings continually every fourth week after enrollment into the programme. However, the mean time between screenings in the data is 8.1 weeks. To accomodate this, it is decided to restrict feature vector 3 to include citizens with a second screening 4-8 weeks after enrolling in the programme. In order to leverage the information in the training data in this interval, feature vector 3 is further restricted to only include citizens with at least two completed training session. This is done as some features such as the mean time between training cannot be calculated for intervals with one or less training sessions. 11 subjects out of the 230 in the feature vector has less than two completed training sessions. These are removed from the data in feature vector 3. ($n = 219$).

4.4.2.3 Feature vector examples

This section aims to provide an example of the rehabilitation of three fictional citizens and how they would be included or not included in the three feature vectors. An illustration of these can be seen in figure 4.6. The first example is of the citizen *Edith*. She

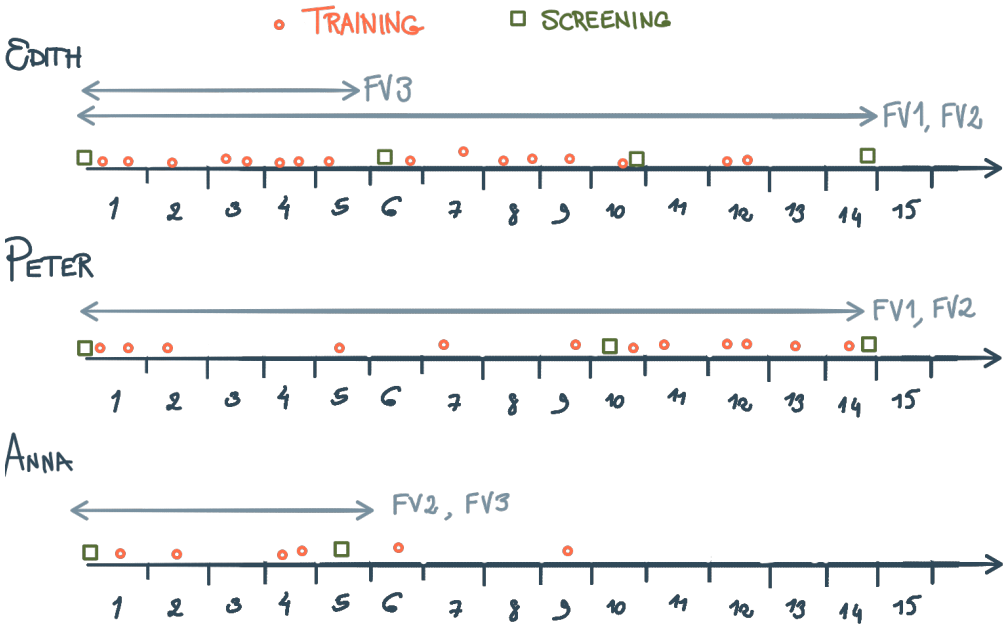


Figure 4.6: *Fictional examples of citizens and how they would or would not be included in the three feature vectors.*

has been training consistently, only missing few training sessions throughout the first 12 weeks. In addition, she has screenings recorded at the start of the programme and again in weeks 6, 10, and 14. This means she would be included in all three feature vectors and thereby also all experiments. For the first and second feature vectors Edith's screening at the beginning of the programme and her screening in week 14 would be used to calculate the development in the *need for help* score. For feature vector 3 the first

screening would be used, and then her screening in week 6 would mark the end of the screening interval, as the interval must be from four to eight weeks.

The second citizen in the example is *Peter*. He started out training fine in the first two weeks, but then he had several weeks of no training and more sporadically completed sessions, but in week nine, he started training consistently until his program ended in week 14. He has an initial screening, and then the next screening is not conducted until week 10, while the final screening is at week 14. Because he has no screenings in week four to eight, he would not be included in feature vector 3, but would be included in feature vectors 1 and 2.

Third is *Anna*. She started training, but then failed to uphold the programme after her last training session in week 9. Since she does not have a screening after week 12, she would not be included in feature vector 1, but would be present in the other feature vector 2 and 3.

4.4.3 Candidate predictors

The establishment of three feature vectors as described in subsection 4.4.2 on page 39 yields different sets of predictors available for training the model. Common to feature vectors 1 and 2 is that the set of known predictors is alike as the objective is to make a prediction based on data available at the initiation of a rehabilitation programme. This sparse set of features includes the age and gender of the citizen as well as scores for *need for help* and *physical strength*.

The impact analysis completed by DigiRehab in 2019 [23] provides a rehabilitation indicator as calculated by the proportion between *need for help* score and *physical strength* score as given in equation 4.1.

$$\text{Rehabilitation potential} = \frac{\textit{need for help}\ \text{score}}{\textit{physical strength}\ \text{score}} \tag{4.1}$$

This leads to a total of six features listed below. A table describing each of the features can be found in appendix B.1.

- Age
- Sex
- NumberATsRunning
- NeedsStart
- PhysicsStart
- RehabIndicator

The number of assistive products currently lent by the citizen is also known. Feature vector 3 extends the feature set by 15 features listed below, based on information gathered from the first couple of weeks after the rehabilitation programme is initiated.

- MeanEvaluation
- StdEvaluation
- MinEvaluation
- MaxEvaluation
- nTrainingPrWeek
- nTrainingPrWeekMax
- nTrainingPrWeekMin
- TimeBetweenTrainingsAvg
- nCancellationsPrWeekAvg
- nCancellationsPrWeekMin
- nCancellationsPrWeekMax
- NeedsEnd
- NeedsDiff
- PhysicsEnd
- PhysicsDiff

Additionally, for all three feature vectors, the assistive devices received by the citizen are known. These devices are identified by a hierarchically structured ISO classification number as described in section 4.2.2.1 on page 32. How these can be used as predictors in a model is discussed in the following.

4.4.3.1 Assistive technology as predictors

An aim of this thesis is to investigate how information about the assistive technology of a citizen can be utilized to improve predictive performance. It is therefore, of interest to investigate how this information is best leveraged in a prediction model. To investigate this, different possible transformations of the device features are considered. These include:

- Dummy variable representation of all possible devices.
  - Binary (has device - does not have device)
  - Numerical (the number currently held of a specific device)
- Clusters of device lending patterns.

- Selection of specific devices based on earlier studies.

Furthermore, as the devices are represented as hierarchically structured ISO-class numbers, it is of interest to investigate what level of granularity provides the most useful information for the classifiers. This will be also be considered for the device feature representations. The different device feature representations are described in the following.

Assistive technologies as dummy variables While some algorithms can handle categorical features, this is unfortunately not the case for logistic regression and the scki-kit learn implementation of random forest. As the amount of distinct assistive technology classification numbers surpasses 2, they are encoded as dummy variables, where the K-level qualitative variable is represented by a vector of K binary variables. This implies that the total amount of features increases by the number of distinct assistive products, that the citizens currently hold. Thus, a citizen will have a set of features named *DevicesUnique_isoclass*, where *isoclass* represents an assistive device ISO class number. If the citizen currently holds the assistive product, the feature is a binary 1 and a 0 otherwise. Information regarding the amount of devices within each category is omitted this way, which is why another approach was formed as well. Instead of defining the dummy variables as binary, they are numerical, related to the number of assistive technologies that the citizen currently possesses within the actual category. These features are named *DevicesCount_isoclass*, where *isoclass* represents an assistive device ISO class number. Both approaches ensure that the classifier is presented with all of the distinct ISO class categories. The granularity of the ISO class numbers can be altered to a more generalized form including only 2, 4 or 6 digits. The drawback of using dummy variables exists in the heavily increase in the total amount of predictors p. When the number of subjects n is exceeded by p it is easier to obtain a useless models without residuals, meaning that the model assessment should be done carefully [41, pp. 243-244]. The memory consumption is another factor for high cardinality categorical features, as they result in many dummy features which are treated independently by the algorithms. Another shortcoming of this approach is the loss of the ISO class hierarchical structure in the numbers. However, no tangible method for preserving the information were identified.

Clusters based on device patterns The purpose of this approach is to define a set of clusters each containing similar patterns and place citizens inside one of these clusters. The preceding work on this data succeeded in identifying 84 different clusters [60] and this is being used to create the cluster center initialisation. As opposed to the approach described in section 4.4.3.1, the clusters are based on patterns in the ISO class numbers of a granularity of 4. This improves on the robustness, as technological development in the assistive technologies - e.g. introducing new devices replacing the old - is omitted. The set of cluster centers is passed to a k-modes algorithm, which is fitted on the data on the assistive devices - for a total of 47,360 different citizens between the year of 1977

and 2018. Each subject is assigned to a cluster based on their history of lent devices and the cluster number is added as a categorical feature for each citizen. However, as only a small subset of all subjects are used for the classification as described in section 4.3 on page 34, the amount of clusters was reduced to retain the frame of reference. Restricting the cluster initialisation method, a total of 36 clusters were obtained. When applied to the citizens in the training data not all of the clusters came to use. Table C.1 on page 101 reflects this for the three feature vectors used in the project objectives. Defining clusters for the citizens might seem to capture more information about the lent assistive technologies as historics are utilised. However, the clusters work as a black box reducing the transparency of predictions as they are not easily explained from their numbers. Additionally, as clustering belongs to the unsupervised field of learning, there exist no direct measure of success [39, p. 487], which means that a neutral performance metric of the clustering cannot be obtained. By looking into the values presented in the aforementioned table, it can be seen that a majority of the total amount of clusters are used by the subsets of observations, which implies that the clusters are not too generalised nor specific. However, it was chosen to disregard the use of clusters in this thesis, as the uncertainties regarding the clusters decreases the aim for transparent predictions.

Devices based on earlier studies (DigiRehab-defined) In 2018, Kommunal Sundhed, a magazine aimed at managers working with health care in municipalities, published a feature [61] describing the results of a pilot project from Aalborg Municipality, conducted by DigiRehab. This study showed that intelligent usage of assistive products improves the rehabilitation by 106% compared to training with randomly selected citizens. In this connection, DigiRehab also studied how much more self-reliant citizens could become by training if the focus was placed on the assistive technology the citizens received. The presented results are shown in table 4.7.

Assistive technology	Improvement in self-sufficiency
Systems with lateral tilt function	13%
All rehabilitation programmes in Aalborg	14%
Raised toilet seat	15%
Shower stool	20%
Raised toilet seat and shower stool	27%
Rollator	29%

Table 4.7: *Percentage-wise improvement in self-reliance after a rehabilitation programme among citizens currently lending a certain device. Based on a study of 87 citizens in Aalborg Municipality [61].*

The frame of reference is the 14% improvement as the average of all rehabilitation programmes in Aalborg Municipality, which is why it would be of interest to place a focus on whether the citizen currently possesses a raised toilet seat, a shower stool, a combination of these two devices or a rollator. Citizens using a system with lateral tilt function did not improve above the frame of reference and this assistive technology is therefore not examined further. The four categorical features are defined as *HasRolla-*

tor, *HasRaisedToiletSeat*, *HasShowerStool*, *HasRaisedToiletSeatAndShowerStool*. It would be preferable to let the model find these patterns by itself, but with a limited amount of observations and a high amount of different assistive devices, this is a method of ensuring simplicity. Furthermore, the transparency of the predictors is increased by this method.

Selection of assistive technology features As the cluster representation is considered less transparent, this thesis will use the following three representations of device features as candidate predictors for the classifiers.

- Dummy variable representation of all possible devices.
  - Binary (has device - does not have device)
  - Numerical (the number currently held of a specific device)
- Selection of specific devices based on earlier studies.

Furthermore, it is of interest to determine the optimal level of granularity for the device features in terms of ISO-class numbers. This will be investigated to define the level of granularity for the binary and numerical dummy variable representations. It will be determined through experiments where the level of granularity is varied between the possible four levels of 2, 4, 6, and 8 digit ISO-class numbers. From this experiment one level of granularity will be selected for the further use of the dummy variable representations of the assistive technology features.

4.4.4 Feature Correlation

When the features have been created a feature correlation matrix can be used to investigate the correlation between the candidate features and the response. This will allow for some insights into whether any feature may on it own provide a good predictor of the response. Furthermore a correlation matrix can also provide insights into whether two predictors have a correlation with each other. This can indicate collinearity which can cause problems with the interpretability of the model if both predictors are used and the coefficients correlated features may take on large values as they cancel each other out [41, pp. 99-101].

Three correlation matrices have been created, one for each feature vector. These can be seen in figures 4.7, 4.8, and 4.9. All the base predictors are included, such as *age*, *sex*, and *NumberATsRunning* (the total number of assistive devices a citizen has had). For the device features only the devices based on earlier studies as presented in section 4.4.3.1 are included. This includes devices such as *rollator* and *shower stool*. The other representations of device features are not included in the correlation matrices for practical reasons as there are more than 100 different features in these. It is worth noting that as the number of observations in the feature vectors vary, the correlations differ across the three matrices.

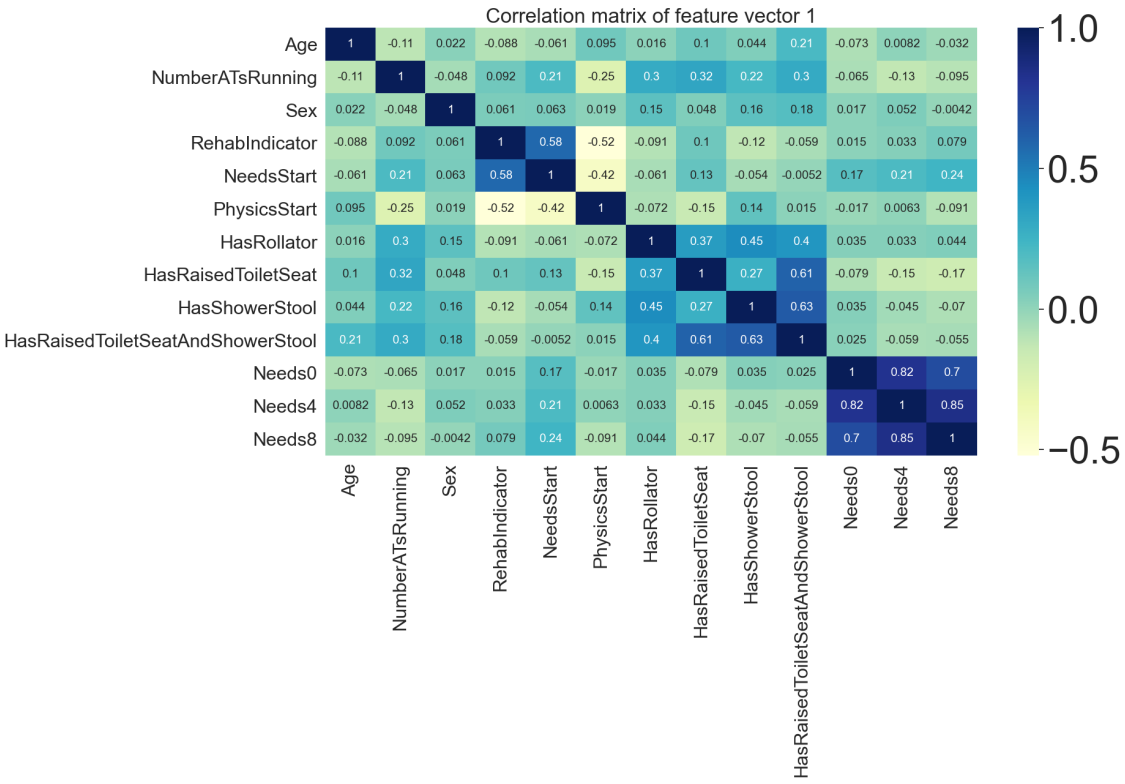


Figure 4.7: *Correlation matrix showing the correlation of the predictors and the response variables of feature vector 1.*

In figure 4.7 the correlation matrix of feature vector 1 can be seen. The features of feature vector 1 are included and so are the response variables *Needs-1*, *Needs-4*, and *Needs-8*. No feature seems to be very highly correlated with the response variables, but *NumberATsRunning*, *NeedsStart* (the need for help at the start of the programme), and *HasRaisedToiletSeat* seem to have the highest (positive or negative) correlation with the response variables. Since no features on their own are highly correlated with the response, it makes sense to investigate the predictive performance when the features are combined in a model.

Some predictors are somewhat correlated with others. This may influence the model. However, for logistic regression, regularisation is applied to alleviate some of challenges with this. The highest feature correlation is seen for *HasRaisedToiletSeatAndShowerStool* and *HasShowerStool*, which makes sense as the do depend on each other. Random forest uses a limited number of features for each tree it grows to alleviate this.

In figure 4.8 the correlation matrix of feature vector 2 is seen. The response variable included is *successfulProgrammeAll*. Again, no features are highly correlated with the response, but *HasRaisedToiletSeatAndShowerStool* seem to have the strongest correlation

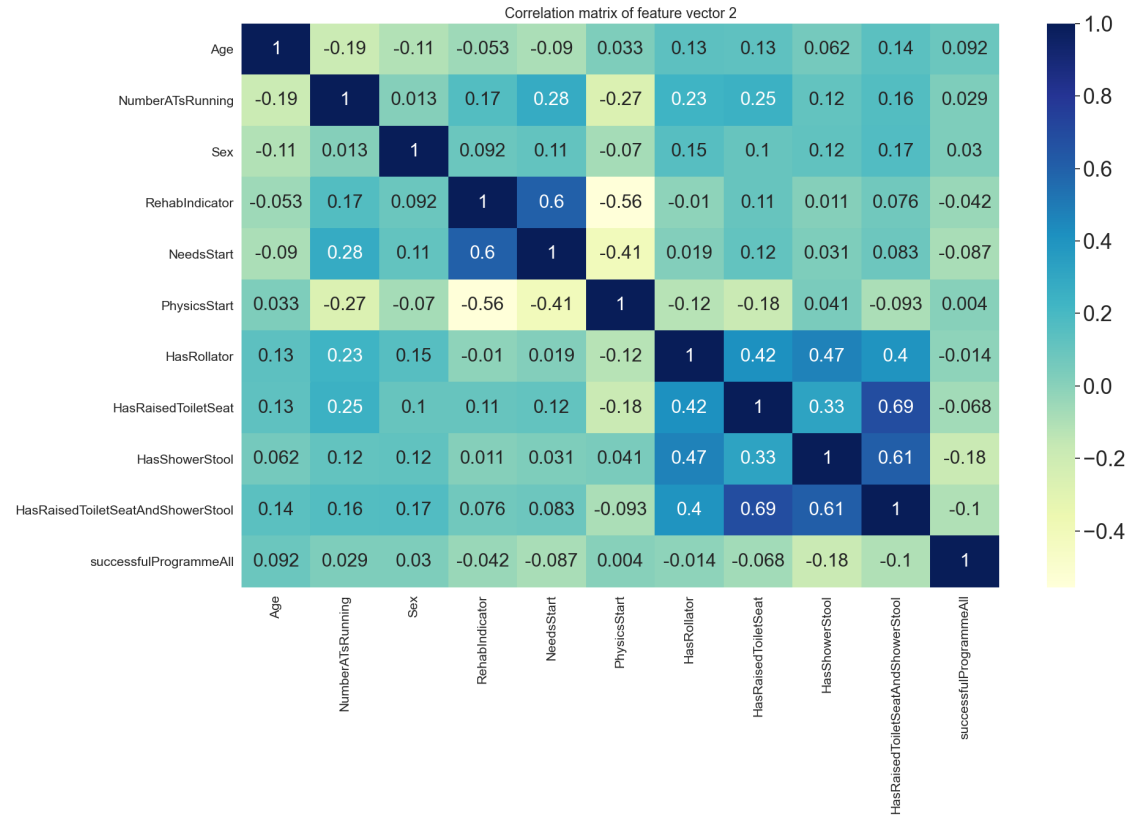


Figure 4.8: *Correlation matrix showing the correlation of the predictors and the response variables of feature vector 2.*

(-0.18), in this case negative. It will therefore also in this case be worth investigating the predictive performance when the features are combined.

Finally, in the correlation matrix of feature vector 3, seen in figure 4.9 additional features are included. These are the features based on information from the first four to eight weeks of training. Some features have quite high correlation values as they may have been based on the same data, like *nCancellationsPrWeekAvg* (average number of cancelled training sessions pr. week) and *nCancellationsPrWeekMin* (the lowest number of cancellations for a week) which are both based on cancellation data regarding the trainings.

The response variable in correlation matrix 3 is again *successfulProgrammeAll*. This has a positive correlation of 0.5 with *nTrainingPrWeek* (the average number of completed training sessions per week in the first 4-8 weeks). This may indicate, that if a citizen started out training consistently, he or she is more likely to complete the program successfully. It is still of interest to investigate how the interaction of these features can perform in predicting the response.

How the features can be combined for best performance will be determined in the

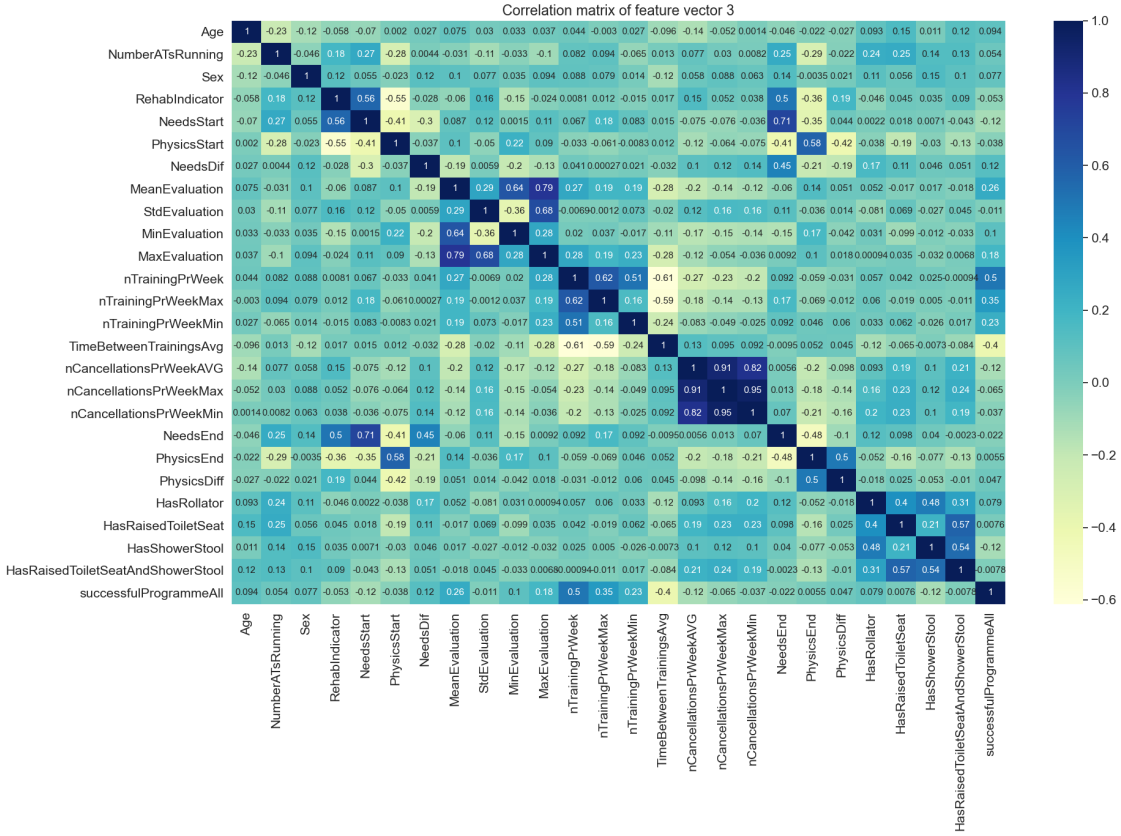


Figure 4.9: *Correlation matrix showing the correlation of the predictors and the response variables of feature vector 3.*

model selection process described in section 5.4.

5

Design and implementation

This chapter outlines the design, implementation and testing along with the environmental set up.

5.1 Conceptual overview

This thesis designs, implements and tests statistical learning algorithms for a reliable and transparent prediction in a citizen's benefit from rehabilitation. The purpose is to map the objectives for the KMD signature project (further described in section 1.3) to an implementation of a clinical decision support system. The conceptual overview is illustrated in figure 5.1.

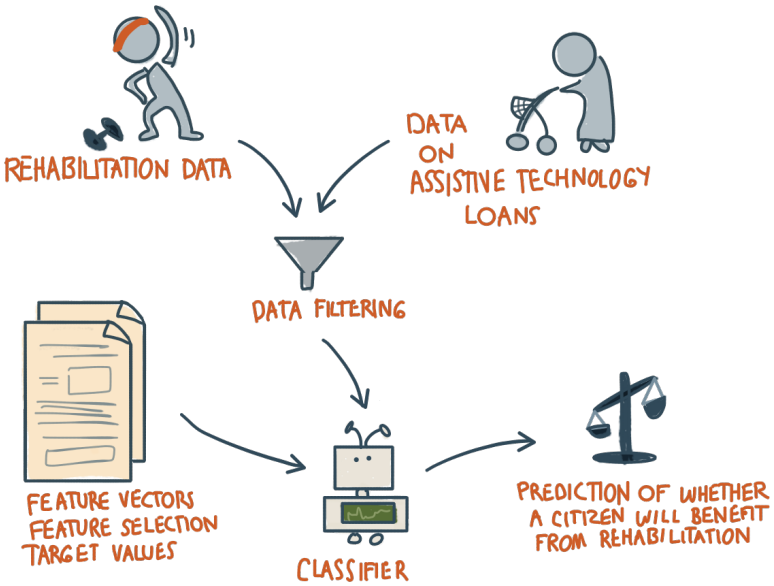


Figure 5.1: *The conceptual overview. The designed and implemented system seeks to achieve the highest AUC score by using state of the art classification algorithms together with feature selection as well as different representations of the feature vector and target value.*

This solution explores various combinations of citizens with various sets of predictors evaluated by various definitions of the target variable. Furthermore, the solution applies two distinct statistical learning models on the data and combines the most important predictors for these. Finally the models are assessed to conclude on their generalisation in prediction capability.

5.2 Experimental environment and tools

This project has been developed in the JetBrains IDE Pycharm [62] using Python 3.7. Python is a popular dynamically-typed language especially used for data science [63]. All releases are open-source. The most important python packages used throughout the development are outlined below.

5.2.1 Scikit-learn

This library provides implementation of a large number of machine learning algorithms [64]. The library is used for the implementation of logistic regression and random forest - and furthermore the stratified Cross-Validation and training/test splitting are also used. In addition to this, Scikit-learn provides metrics for evaluation of the models.

5.2.2 Pandas

Pandas eases the work with complex data structures [65]. This library is used for all the data handling, from loading the data, analysing, altering, sorting it and for the creation of dummy variables.

5.2.3 LIME

LIME is the library used for explaining predictions of a machine learning classifier [66]. It supports explanations for textual data as well as tabular data and images. As the data in this thesis is numerical and categorical, the tabular explanations are used.

5.3 Design and implementation overview

This section outlines the design and implementation of the experimental setup used in this thesis. An illustration depicting this is seen in figure 5.2. Firstly, the data is loaded and cleaned. This is elaborated upon in section 4.3. The data is then partitioned into multiple files. One partition of the data is used for feature selection to yield the best performing subset for the model based on the AUC. The other partition of the data is used for model assessment of the best performing models. The next section describes the data filtering in detail.

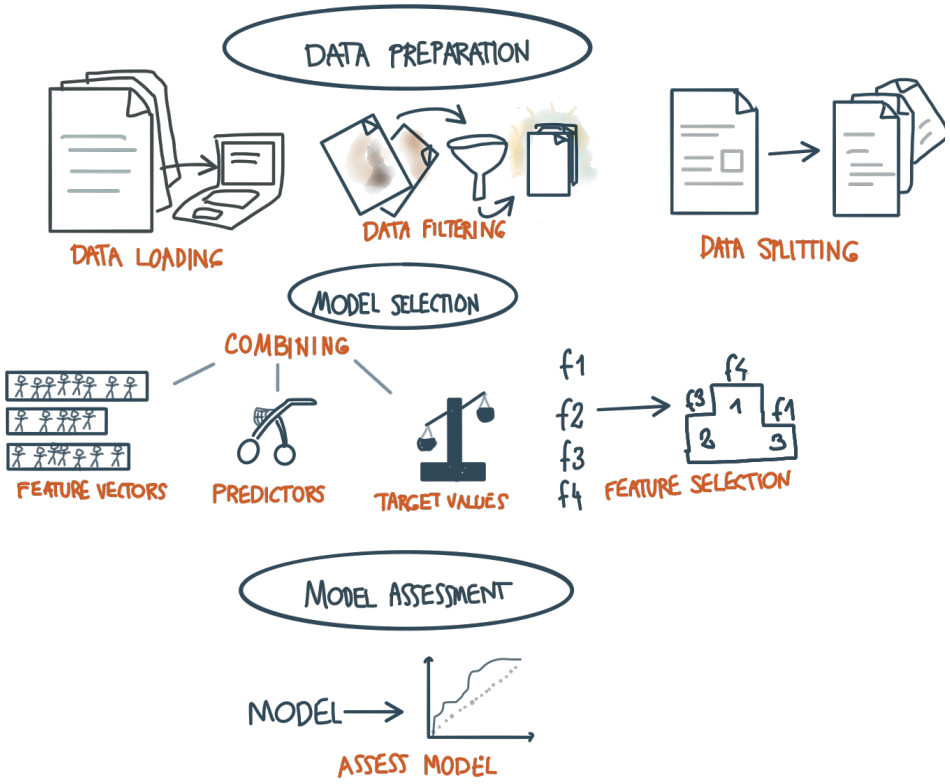


Figure 5.2: *The implementation overview.*

5.4 Model selection and assessment

This section outlines the process of selecting and assessing the models. An illustration of the process of model selection and assessment for the experiments are seen in figure 5.3 on the next page.

5.4.1 Model selection

After loading, wrangling and creating the dummy variables, the feature vectors are partitioned into a training set and a validation set as motivated in section 3.4. The training set consists of 80% of the observations, while the other 20% make up the validation set. The partitions are stratified such that they contain approximately the same proportion of labels in both of the sets. For all experiments a 5-fold Cross-Validation was used. The data is split into 5 partitions of equal size. This results in five experiments and the model performance is calculated as the mean of the AUC score on each of the five test splits. The number of subjects in feature vector 1 is as low as 125. Furthermore, the labels might be skewed according to the threshold of *need for help* improvement and ordinary k-fold Cross-Validation might create splits only containing labels from one of the classes. Therefore, the splits were stratified according to the labels. The model selection

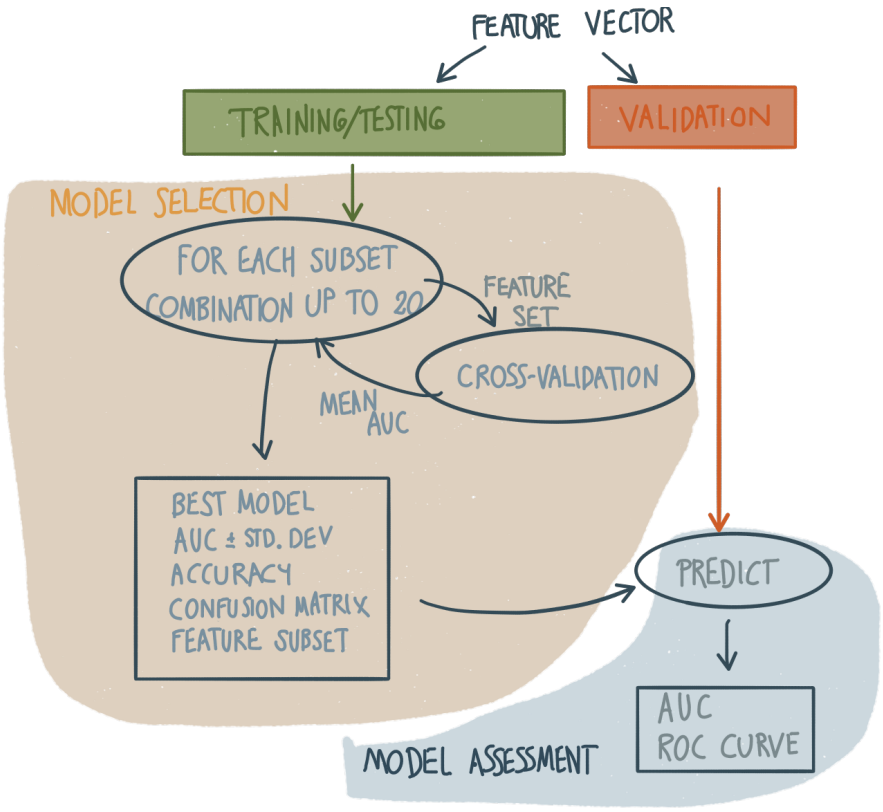


Figure 5.3: *The model selection and assessment of an experiment.*

is concluded when all of the combinations of subsets generated by forwards stepwise subset selection, with a limit of 20 features, has been tested. The resulting best model is saved with information regarding the AUC and standard deviation, the accuracy, the averaged and normalised confusion matrix and the feature subset.

5.4.2 Model assessment

The model assessment is the final step of the experimental setup. The best model obtained from the model selection is trained on the entire training set, and now it must predict on the validation set which is data it has not encountered in the fitting process. This provides the measure of generalisation prediction capability. The AUC is calculated and a ROC curve is generated, which can be used to assess the model.

5.5 Overview of the experiments

This sections outlines how the experiments are structured based on the data preparation described in section 4.3. An overview of all five experiments is given in figure 5.4 on the facing page. The aim of this thesis is two-fold and the conducted experiments

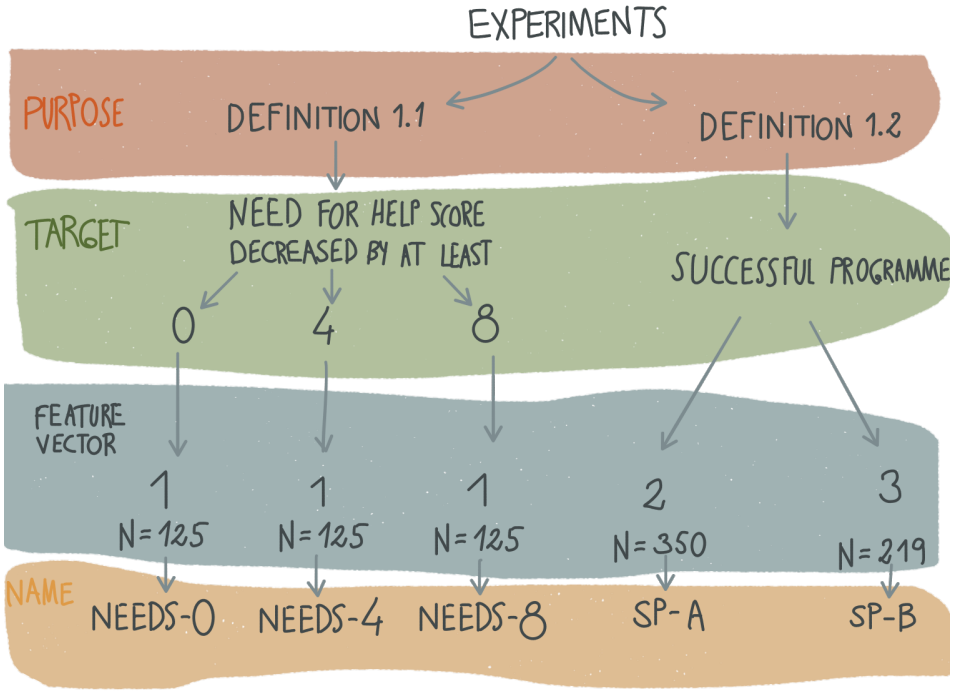


Figure 5.4: *Overview of the five experiments.*

reflect this.

Firstly, it is investigated which home care patients will benefit from exercise, where the benefit is evaluated by the development in a *need for help* score as defined in definition 1.1 on page 4. For the models addressing this objective, the predictors are constructed from data available prior to enrollment in the rehabilitation programme from DigiRehab.

For the second objective it is investigated which home care patients are likely to complete the DigiRehab exercise program successfully as defined in definition 1.2 on page 4. This branches into two experiments. The first is based on the predictors available prior to enrollment in the rehabilitation programme, as with the predictions of *need for help* score. The other experiment is based on the data available from the first two screenings in the programme. This allows the models to utilize information gathered from the first weeks of exercise after enrollment into the program to predict whether the citizen will complete the rest of the programme. For both project objectives the predictions are evaluated based on the AUC metric as described in section 3.3.

6

Experiments and results

This chapter describes the five different experiments and the results obtained from the various approaches described in the preceding chapter. The chapter is partitioned into three sections; one describing the model selection results, one for the model assessment results and one for the results of applying LIME on the models to obtain explanations.

6.1 Model selection

This section outlines the results from the trained models during model selection. Two objectives are studied; the optimum amount of granularity in the assistive technology ISO class, described in section 6.1.1 and the optimum combination of features in section 6.1.3 and 6.1.4. In section 6.1.2, a list describing the categories for assistive technologies is shown.

6.1.1 Granularity of assistive device ISO classes

Initially, it is investigated which level of granularity in the assistive technology ISO-classes in general provides the best predictions. For each classification algorithm, all four levels of granularity were examined; 2, 4, 6 and 8 digits. A higher number of digits corresponds to a more specific description of the assistive technology. This is relevant when classifying assistive devices as categorical data as described in section 4.4.3.1. The mean AUC was calculated for each of the five experiments based on predicting with the *DevicesUnique* and *DevicesCount* predictors.

6.1.1.1 Experiment GRAN-A: Granularity of ISO classes for the logistic regression classifier

Purpose Examining what level of granularity for the device ISO classes yields the highest AUC score for the logistic regression classifier.

Data and model This experiment considers all five experiments, but only evaluates on the AUC for predictions based on the *DevicesUnique* and the *DevicesCount* predictors. The model used is logistic regression.

Results In figure 6.1 the AUC scores for logistic regression with varying granularity of ISO classes are plotted for each of the five experiments. The mean for each level of granularity is shown as well. It can be seen that logistic regression obtains a higher averaged AUC score when the ISO class numbers are six digits.

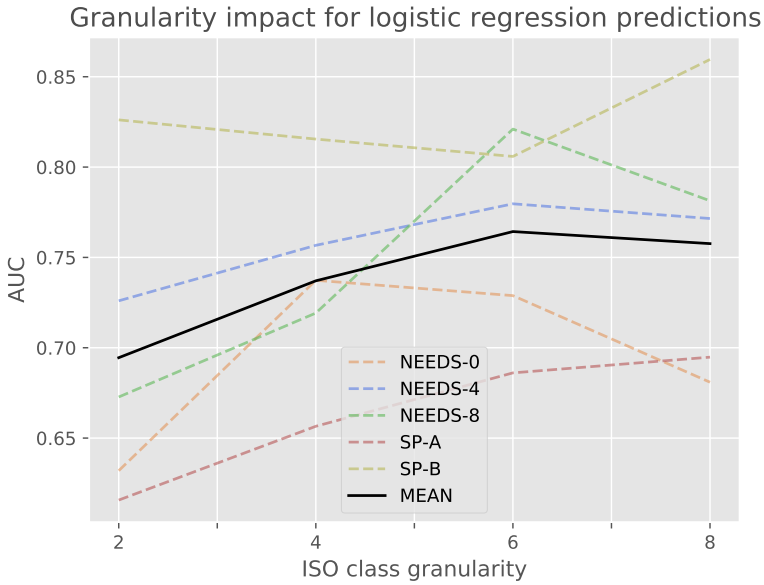


Figure 6.1: *The impact of altering the granularity of the device ISO classes in regards to prediction performance for logistic regression. Six digits yield the highest mean AUC.*

The results for the plot are presented in table 6.1.

Experiment	2 digits	4 digits	6 digits	8 digits
NEEDS-0	0.632	0.737	0.729	0.681
NEEDS-4	0.726	0.757	0.780	0.772
NEEDS-8	0.673	0.719	0.821	0.781
SP-A	0.616	0.657	0.686	0.695
SP-B	0.826	0.815	0.806	0.860
MEAN	0.694	0.737	0.764	0.758

Table 6.1: *Mean values of AUC computed for logistic regression predictions on the basis of DevicesUnique and DevicesCount for each of the five experiments described in section 6.1.3 and 6.1.4*

Discussion Based on the results, it can be seen that logistic regression yields the highest mean AUC score when the device ISO classes have a length of six digits. However, it varies depending on the experiment. Both experiments with the successful programme as target value (SP-A and SP-B) perform best with eight digit ISO classes using logistic regression. The mean AUC does not vary much and therefore it would be best to apply the granularity individually for each experiment. However, applying the best-fit ISO

class granularity for each experiment will increase the complexity of the subsequent experiments and reduce the frame of reference between each experiment. Therefore, the mean AUC is used to determine the granularity of 6 digits.

6.1.1.2 Experiment GRAN-B: Granularity of ISO classes for the random forest classifier

Purpose Examining what level of granularity for the device ISO classes yields the highest AUC score for the random forest classifier.

Data and model This experiment considers all five experiments, but only evaluates on the AUC for predictions based on the *DevicesUnique* and the *DevicesCount* predictors. The model used is random forest.

Results In figure 6.2 the AUC scores for random forest with varying granularity of ISO classes are plotted for each of the five experiments. The mean for each level of granularity is shown as well.

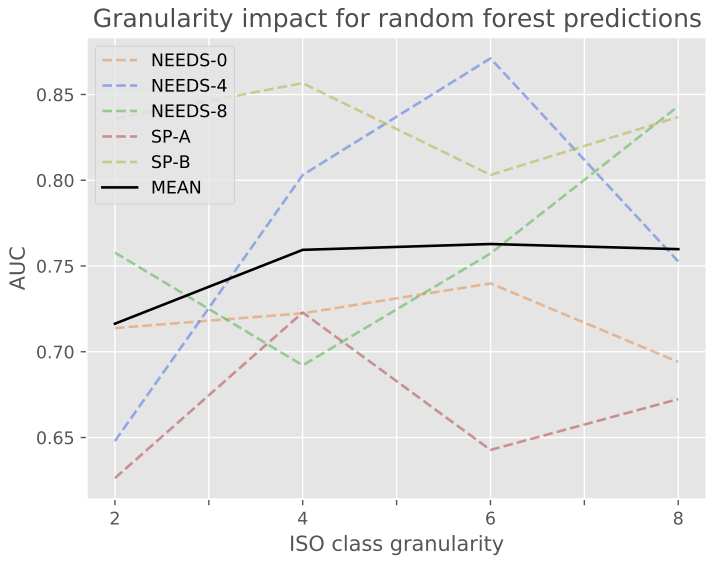


Figure 6.2: *The impact of altering the granularity of the device ISO classes in regards to prediction performance for random forest. Six digits yield the highest mean AUC.*

The results for the plot are presented in table 6.2 on the facing page.

Experiment	2 digits	4 digits	6 digits	8 digits
NEEDS-0	0.714	0.722	0.740	0.694
NEEDS-4	0.648	0.803	0.871	0.753
NEEDS-8	0.758	0.692	0.757	0.843
SP-A	0.626	0.723	0.643	0.672
SP-B	0.836	0.857	0.803	0.837
MEAN	0.716	0.759	0.763	0.760

Table 6.2: *Mean values of AUC computed for random forest predictions on the basis of DevicesUnique and DevicesCount for each of the five experiments described in section 6.1.3 and 6.1.4*

Discussion Based on the results, it can be seen that random forest, when looking at the mean, yields the highest AUC score when the device ISO classes have a length of six digits. However, it varies a lot depending on the experiment. Both experiments with the successful programme as target value (SP-A and SP-B) perform best with ISO classes of four digits. The mean AUC does not vary much and therefore it would be best to apply the granularity individually for each experiment. However, applying the best-fit ISO class granularity for each experiment will increase the complexity of the subsequent experiments and reduce the frame of reference between each experiment. Therefore, the mean AUC is used to determine the granularity of 6 digits.

6.1.2 Description of the assistive technology categories

The assistive aids are divided into eight main categories listed below with a colour representing each category and a description of the product series within each category. The categories are defined by The National Board of Social Services [56].

04:Assistive products for measuring, supporting, training or replacing body functions
Products that monitor or assess a persons medical condition, and products that support, or provide a substitute for, a specific body function. Included examples are products used in medical treatment. Excluded are assistive products used exclusively by healthcare professionals.

09: Assistive products for self-care activities and participation in self-care
Assistive products for toileting, incontinence, personal hygiene, sexual activities, etc. Included are also clothes.

12: Assistive products for personal mobility and transportation
Products intended to support or replace a persons capacity to move indoors and outdoors, to transfer from one place to another or to use personal or public transportation. Walking aids, wheelchairs, cycles, vehicle adaptations, assistive products for transfer and turning, for lifting persons, and for orientation.

15: Assistive products for domestic activities and participation in domestic life
Assistive products for cooking, dishwashing, eating and drinking, cleaning, maintaining textiles etc.

18: Furnishings, fixtures and other assistive products for supporting activities in indoor and outdoor human-made environments
Lighting, tables, chairs, seats and cushions, beds and mattresses, grab bars, lifting platforms, stairlifts, ramps, etc.

22: Assistive products for communication and information management
Assistive products for seeing, hearing, speech, writing, reading, calculation, recording and playing sound, telephoning, alarming, and the use of information technology.

24: Assistive products for controlling, carrying, moving and handling objects and devices
Packaging openers, grip adapters, assistive products for operating and controlling, for grasping, for fixation, and for carrying and transporting objects.

These colour codes are used throughout section 6.1.3 and 6.1.4 to ease the distinction of main categories for the assistive products.

6.1.3 Predicting development in the *need for help* score

By the definition given in 1.1, the aim of the succeeding experiments within this section is to predict a citizen's rehabilitation benefit based on the *need for help* score. The development in the *need for help* score is evaluated by three thresholds. All results are divided into three subsections below, each describing one of the threshold.

Only citizens which have enrolled in a rehabilitation programme of at least 12 weeks are evaluated upon. Further description of the choice of feature vector 1 is found in section 4.4.2 on page 39. The feature vector is identical for all three experiments.

6.1.3.1 Experiment NEEDS-0: Predicting a decrease in the *need for help* score by at least 0

Purpose The purpose of this experiment is to find the model and the feature subset yielding the best prediction. The target value is defined as whether the *need for help* score after the rehabilitation programme is equal to or lower than the score at the first screening.

Data and model The data is based on feature vector 1 containing 125 subjects as described in section 4.4.2. Out of a total of 125 citizens, 67 have had benefit from the rehabilitation programme while 58 have not. The distribution is 53.6% and 46.4%. This response value is evaluated using logistic regression and random forest classification methods.

Results Resulting AUCs are listed in table 6.3

	Logistic regression		**Random forest**	
AT predictors	AUC	ACC	AUC	ACC
None	0.603 ($\pm$0.099)	0.560	0.662 ($\pm$0.059)	0.600
DigiRehab-defined[1]	0.603 ($\pm$0.099)	0.560	0.717 ($\pm$0.12)	0.690
DevicesUnique	**0.732** ($\pm$0.079)	0.600	**0.743** ($\pm$0.14)	0.710
DevicesCount	0.727 ($\pm$0.083)	0.620	0.736 ($\pm$0.11)	0.680

Table 6.3: *Results for predicting a stagnated or decreased need for help score using logistic regression and random forest.* [1] *i.e. HasRollator*

The logistic regression yields an AUC of 0.732 ($\pm$0.079) whilst random forest predicts with an AUC of 0.743 ($\pm$0.14), both using the *DevicesUnique* predictors for the assistive technology currently lent by the citizen. The normalised confusion matrices for the best performing logistic regression and random forest classifier are illustrated in figure 6.3 on the next page.

	p	n
p′	0.589	0.385
n′	0.411	0.615

(a) *Logistic regression*

	p	n
p′	0.611	0.205
n′	0.389	0.795

(b) *Random forest*

Figure 6.3: *Confusion matrices for predicting an equal or decreased need for help score for logistic regression (a) and random forest (b).*

Logistic regression used a set of 20 features to yield the highest AUC. In table 6.4 the features are listed by importance and marked with the color of the respective category.

Rank	Feature	Description
1	NeedsStart	-
2	120316	Multi-tip walking sticks and canes
3	Age	-
4	181003	Back supports
5	180903	Chairs
6	222718	Personal emergency alarm systems
7	120306	Elbow crutches
8	090903	Assistive products for putting on or removing socks and pantyhose
9	091209	Toilet seats
10	120727	Products to hold assistive products for walking in place when not in use
11	221824	FM systems
12	221830	Telecoils
13	181233	Bed extenders
14	222403	Standard network telephones
15	222704	Signalling devices
16	220309	Magnifier glasses, lenses and lens systems for magnification
17	123612	Stationary hoists fixed to walls, floor or ceiling
18	181210	Beds and detachable bed boards/mattress support platforms with powered adjustment
19	093304	Bath boards
20	180907	Standing chairs

Table 6.4: *Feature subset used for predicting a decrease in the need for help score by at least 0 using the logistic regression classifier.*

Random forest used a set of 18 features to ensure the highest AUC. In table 6.5, the

features are listed by importance and marked with the color of the respective category.

Rank	Feature	Description
1	NeedsStart	-
2	Age	-
3	180903	Chairs
4	181006	Seat cushions and underlays
5	091218	Raised toilet seats fixed to toilet
6	222704	Signalling devices
7	181003	Back supports
8	091233	Bedpans
9	123106	Turntables
10	120316	Multi-tip walking sticks and canes
11	120606	Rollators
12	122439	Devices to which a wheelchair is attached that facilitate movement up and down stairs
13	122203	Bimanual handrim-drive wheelchairs
14	220309	Magnifier glasses, lenses and lens systems for magnification
15	123604	Mobile hoists for transferring a person in standing position
16	120306	Elbow crutches
17	222712	Clocks and timepieces
18	091209	Toilet seats

Table 6.5: *Feature subset used for predicting a stagnated or decreased need for help score using the random forest classifier.*

Discussion This experiment shows the importance of including assistive device predictors in the feature set. The best-performing models used the same feature category *DevicesUnique* representing the devices to predict, but the final feature set differed for both models. Only devices from four different categories were selected. For logistic regression, 20 features were used. This implies that the model might be able to perform better if more than 20 features were allowed.

6.1.3.2 Experiment NEEDS-4: Predicting an improvement in *need for help* score by at least 4

Purpose The purpose of this experiment is to find the model and the feature subset yielding the best prediction. The target value is defined as whether the *need for help* score after the rehabilitation programme has decreased by at least 4 compared to the score at the first screening.

Data and models The data is based on feature vector 1 containing 125 subjects as described in section 4.4.2. Out of a total of 125 citizens, 55 have had benefit from the rehabilitation programme while 70 have not. Thus the distribution is 44% and 56% respectively. This response value is evaluated using logistic regression and random forest classification methods.

Results Resulting AUCs are listed in table 6.6

	Logistic regression		**Random forest**	
AT predictors	AUC	ACC	AUC	ACC
None	0.648 ($\pm$0.13)	0.550	0.763 ($\pm$0.011)	0.700
DigiRehab-defined[1]	0.707 ($\pm$0.097)	0.610	0.823 ($\pm$0.093)	0.720
DevicesUnique	**0.783** ($\pm$0.073)	0.630	**0.878** ($\pm$0.086)	0.750
DevicesCount	0.776 ($\pm$0.077)	0.640	0.864 ($\pm$0.057)	0.740

Table 6.6: *Results for predicting a decrease in the need for help score of at least 4 using logistic regression and random forest.* [1] *i.e. HasRollator*

The logistic regression classifier yields an AUC of 0.783 ($\pm$0.073) using the *DevicesUnique* predictors for the assistive technology possessed by the citizen, whilst random forest classification obtains an AUC of 0.878 ($\pm$0.086), also using the *DevicesUnique* as assistive device predictors. The normalised confusion matrices for the best performing logistic regression and random forest classifier are illustrated in figure 6.4.

	p	**n**
p′	0.585	0.322
n′	0.415	0.678

(a) *Logistic regression*

	p	**n**
p′	0.861	0.381
n′	0.139	0.619

(b) *Random forest*

Figure 6.4: *Confusion matrices for predicting a decrease in the need for help score of at least 4 for logistic regression (a) and random forest (b).*

Logistic regression used a set of 15 features to yield its best performance. In table 6.7, the features are listed by importance.

Rank	Feature	Description
1	NeedsStart	-
2	180903	Chairs
3	122218	Push wheelchairs
4	120316	Multi-tip walking sticks and canes
5	220309	Magnifier glasses, lenses and lens systems for magnification
6	122203	Bimanual handrim-drive wheelchairs
7	242103	Manual gripping tongs
8	NumberATsRunning	-
9	120606	Rollators
10	123603	Mobile hoists for transferring a person in sitting position with sling seats
11	180907	Standing chairs
12	091209	Toilet seats
13	222712	Clocks and timepieces
14	180315	Bed tables
15	122439	Devices to which a wheelchair is attached that facilitate movement up and down stairs

Table 6.7: *Feature subset used for predicting a decrease in the need for help score of at least 4 using the logistic regression classifier.*

The forwards subset selection algorithm combined 19 features to ensure the highest AUC for random forest. In table 6.8 on the next page, the features are listed by importance.

Rank	Feature	Description
1	RehabIndicator	-
2	NumberATsRunning	-
3	043303	Seat cushions and underlays for tissue integrity
4	242103	Manual gripping tongs
5	180907	Standing chairs
6	091218	Raised toilet seats fixed to toilet
7	043306	Assistive products for tissue integrity when lying down
8	220309	Magnifier glasses, lenses and lens systems for magnification
9	091233	Bedpans
10	093304	Bath boards
11	090903	Assistive products for putting on or removing socks and pantyhose
12	120727	Products to hold assistive products for walking in place when not in use
13	120606	Rollators
14	222718	Personal emergency alarm systems
15	090706	Positioning pillows, positioning cushions and positioning systems
16	181218	Mattresses and mattress coverings
17	123604	Mobile hoists for transferring a person in standing position
18	180903	Chairs
19	181503	Leg extenders

Table 6.8: *Feature subset used for predicting a decrease in the need for help score of at least* 4 *using the random forest classifier.*

Discussion In this experiment, the random forest classifier clearly outperformed logistic regression in terms of the AUC. In the confusion matrices depicted in figure 6.4 on page 62, it can be seen that random forest is especially good in predicting true positives, while logistic regression is better in predicting true negatives. This could indicate that an ensemble of both might perform even better, though at the cost of higher complexity and lower transparency. The best-performing models used the same feature category *DevicesUnique* representing the devices to predict, but the final feature set differed a lot for both models.

6.1.3.3 Experiment NEEDS-8: Predicting an improvement in *need for help* score by at least 8

Purpose The purpose of this experiment is to find the model and the feature subset yielding the best prediction. The target value is defined as whether the *need for help* score after the rehabilitation programme has improved by at least 8 compared to the score at the first screening.

Data and models The data is based on feature vector 1 containing 125 subjects as described in section 4.4.2. Out of a total of 125 citizens, 45 have had benefit from the rehabilitation programme while 80 have not. The distribution is 36% and 64%. This response value is evaluated using logistic regression and random forest classification methods.

Results Resulting AUCs are listed in table 6.9

	Logistic regression		**Random forest**	
AT predictors	AUC	ACC	AUC	ACC
None	0.682 ($\pm$0.065)	0.610	0.664 ($\pm$0.063)	0.630
DigiRehab-defined[1]	0.766 ($\pm$0.061)	0.620	0.730 ($\pm$0.094)	0.680
DevicesUnique	0.815 ($\pm$0.027)	0.680	**0.759** ($\pm$0.076)	0.670
DevicesCount	**0.827** ($\pm$0.030)	0.720	0.756 ($\pm$0.087)	0.680

Table 6.9: *Results for predicting a decrease in the need for help score of at least 8 using logistic regression and random forest.*[1] *i.e. HasRollator*

The logistic regression yields an AUC of 0.827 ($\pm$0.030) using the *DevicesCount* predictors for the assistive technology possessed by the citizen, whilst random forest classification obtains an AUC of 0.756 ($\pm$0.087), using the *DevicesUnique* as assistive device predictors. The normalised confusion matrices for the best performing logistic regression and random forest classifier are illustrated in figure 6.5.

	p	**n**
p′	0.733	0.307
n′	0.267	0.693

(a) *Logistic regression*

	p	**n**
p′	0.779	0.532
n′	0.221	0.468

(b) *Random forest*

Figure 6.5: *Confusion matrices for predicting a decrease in the need for help score of at least 8 for logistic regression (a) and random forest (b).*

Logistic regression used a set of 15 features to yield its best performance. In table 6.10 on the next page, the features are listed by importance.

Rank	Feature	Description
1	NeedsStart	-
2	091218	Raised toilet seats fixed to toilet
3	123106	Turntables
4	122218	Push wheelchairs
5	180903	Chairs
6	120606	Rollators
7	091233	Bedpans
8	220309	Magnifier glasses, lenses and lens systems for magnification
9	181503	Leg extenders
10	043306	Assistive products for tissue integrity when lying down
11	180315	Bed tables
12	123603	Mobile hoists for transferring a person in sitting position with sling seats
13	Sex	-
14	NumberATsRunning	
15	120306	Elbow crutches

Table 6.10: *Feature subset used for predicting a decrease in the need for help score of at least* 8 *using the logistic regression classifier.*

The forwards subset selection algorithm combined 15 features to ensure the highest AUC for random forest. In table 6.11 on the facing page, the features are listed by importance.

Rank	Feature	Description
1	RehabIndicator	-
2	091203	Commode chairs
3	091218	Raised toilet seats fixed to toilet
4	181218	Mattresses and mattress coverings
5	122218	Push wheelchairs
6	180315	Bed tables
7	123612	Stationary hoists fixed to walls, floor or ceiling
8	093304	Bath boards
9	090903	Assistive products for putting on or removing socks and pantyhose
10	123604	Mobile hoists for transferring a person in standing position
11	181503	Leg extenders
12	222712	Clocks and timepieces
13	123109	Not mounted rails for self-lifting
14	091233	Bedpans
15	123603	Mobile hoists for transferring a person in sitting position with sling seats

Table 6.11: *Feature subset used for predicting a decrease in the need for help score of at least 8 using the random forest classifier.*

Discussion This is the first experiment where logistic regression outperforms random forest in terms of AUC. It is particularly seen in the confusion matrices illustrated in figure 6.5 on page 65, where logistic regression predicts both true positives and negatives fairly well compared to random forest which does not predict true negatives as well as true positives. Even though positives in this experiment is the minority class represented in 36% of the data, the models generally predict well in true negatives. Both of the selected feature sets were of equal size, but they differed in the features with no specific patterns. Additionally, compared to the NEEDS-0 and NEEDS-4 experiments, this has selected the lowest number of features so far.

6.1.4 Predicting who will complete a training programme

DigiRehab has defined a successful rehabilitation programme as having completed training sessions in 8 out of 12 weeks. This is another measure of whether the citizen benefits from training as stated in definition 1.2. Two experiments were conducted - one where the set of features are similar to the ones used to predict the development in *need for help* score and one where the first two screenings were used. Both response values are the same, the difference lies in the predictor features available and the feature vector.

6.1.4.1 Experiment SP-A: Predicting whether the citizen completes a rehabilitation programme based on the first screening

Purpose The purpose of this experiment is to find the model and the feature subset yielding the best prediction. The target value is defined as whether the citizen has completed a rehabilitation programme.

Data and models The data is based on feature vector 2 containing 350 subjects as described in section 4.4.2. This includes all citizens which have had a rehabilitation programme of at least 4 weeks. In total 350 citizens of which 221 have completed a programme and 129 have not. The distribution is 63.1% and 36.9%. This response value is evaluated using logistic regression and random forest classification methods.

Results Resulting AUCs are listed in table 6.12

	Logistic regression		**Random forest**	
AT predictors	AUC	ACC	AUC	ACC
None	0.561 ($\pm$0.083)	0.554	0.563 ($\pm$0.10)	0.550
DigiRehab-defined[1]	0.627 ($\pm$0.069)	0.614	0.606 ($\pm$0.10)	0.600
DevicesUnique	**0.687** ($\pm$0.084)	0.650	**0.654** ($\pm$0.10)	0.604
DevicesCount	0.685 ($\pm$0.064)	0.661	0.632 ($\pm$0.10)	0.561

Table 6.12: *Results for predicting a successful programme based on the first screening using logistic regression and random forest.* [1] *i.e. HasRollator*

The logistic regression yields an AUC of 0.687 ($\pm$0.084) using the *DevicesUnique* predictors for the assistive technology possessed by the citizen, whilst random forest classification obtains an AUC of 0.654 ($\pm$0.10), also using the *DevicesUnique* as assistive device predictors. The normalised confusion matrices for the best performing logistic regression and random forest classifier are illustrated in figure 6.6 on the next page.

	p	n
p′	0.484	0.254
n′	0.516	0.746

(a) *Logistic regression*

	p	n
p′	0.445	0.306
n′	0.555	0.694

(b) *Random forest*

Figure 6.6: *Confusion matrices for predicting a successful programme based on the first screening for logistic regression (a) and random forest (b).*

Logistic regression used a set of 12 features to yield its best performance. In table 6.13, the features are listed by importance.

Rank	Feature	Description
1	093307	Shower chairs with and without wheels
2	NeedsStart	-
3	123109	Not mounted rails for self-lifting
4	222718	Personal emergency alarm systems
5	181503	Leg extenders
6	090903	Assistive products for putting on or removing socks and pantyhose
7	Sex	-
8	091203	Commode chairs
9	150303	Assistive products for weighing and measuring to prepare food and drink
10	181006	Seat cushions and underlays
11	122442	Devices attached to wheelchairs to hold or carry objects
12	122439	Devices to which a wheelchair is attached that facilitate movement up and down stairs

Table 6.13: *Feature subset used for predicting a successful programme based on the first screening using the logistic regression classifier.*

The forwards subset selection algorithm combined 10 features to ensure the highest AUC for random forest. In table 6.14 on the next page, the features are listed by importance.

Rank	Feature	Description
1	093307	Shower chairs with and without wheels
2	Sex	-
3	091203	Commode chairs
4	123604	Mobile hoists for transferring a person in standing position
5	221830	Telecoil amplifiers
6	220309	Magnifier glasses, lenses and lens systems for magnification
7	093304	Bath boards
8	221824	FM systems
9	091215	Toilet seat inserts
10	122439	Devices to which a wheelchair is attached that facilitate movement up and down stairs

Table 6.14: *Feature subset used for predicting a successful programme based on the first screening using the random forest classifier.*

Discussion This experiment yields two low AUC scores, and both models predicted true negatives best while having a hard time even classifying half of the true positives correctly. Both models agree on using the *DevicesUnique* features with a feature set of only 10 and 12 respectively. The gender of the citizen exists in both feature sets and *Shower chairs with and without wheels* is the most important feature for both models.

6.1.4.2 Experiment SP-B: Predicting whether the citizen completes a rehabilitation programme based on the first two screenings

Purpose The purpose of this experiment is to find the model and the feature subset yielding the best prediction. The target value is defined as whether the citizen has completed a rehabiliation programme.

6.1.5 Prediction based on the first two screenings

In this experiment it was examined whether the results obtained at the second screening as well could increase the prediction accuracy. The set of features for this experiment is therefore extended compared to the other experiments as described in section 4.4.3.

Data and models The data is based on feature vector 3 containing 219 subjects as described in section 4.4.2. This includes all citizens which have had a rehabilitation programme with a second screening between 4 and 8 weeks after the initial screening. In total 219 citizens of which 149 have completed a programme and 70 have not. The distribution is 68.0% and 32.0%. The response value is evaluated using logistic regression and random forest classification methods.

Results Resulting AUCs are listed in table 6.15

	Logistic regression		**Random forest**	
AT predictors	AUC	ACC	AUC	ACC
None	0.778 ($\pm$0.079)	0.680	0.845 ($\pm$0.062)	0.783
DigiRehab-defined[1]	0.793 ($\pm$0.083)	0.720	**0.863** ($\pm$0.059)	0.749
DevicesUnique	**0.814** ($\pm$0.086)	0.726	0.808 ($\pm$0.048)	0.731
DevicesCount	0.798 ($\pm$0.092)	0.674	0.798 ($\pm$0.058)	0.709

Table 6.15: *Results for predicting a successful programme based on the first two screenings using logistic regression and random forest.* [1] *i.e. HasRollator*

The logistic regression yields an AUC of 0.814 ($\pm$0.086) using the DevicesUnique predictors for the assistive technology possessed by the citizen, whilst random forest classification obtains an AUC of 0.863 ($\pm$0.059) using the assistive technology predictors defined by DigiRehab as described in section 4.4.3 on page 41. The normalised confusion matrices for the best performing logistic regression and random forest classifier are illustrated in figure 6.7 on the following page.

	p	n
p′	0.679	0.252
n′	0.321	0.748

(a) *Logistic regression*

	p	n
p′	0.432	0.100
n′	0.568	0.900

(b) *Random forest*

Figure 6.7: *Confusion matrices for predicting a successful programme based on the first two screenings for logistic regression (a) and random forest (b).*

Logistic regression used a set of 13 features to yield its best performance. In table 6.16, the features are listed by importance.

Rank	Feature	Description
1	nTrainingPrWeek	-
2	181503	Leg extenders
3	222718	Personal emergency alarm systems
4	NumberATsRunning	-
5	123109	Not mounted rails for self-lifting
6	221830	Telecoil amplifiers
7	TimeBetweenTrainingsAvg	-
8	093307	Shower chairs with and without wheels
9	181226	Side rails to be fixed to beds
10	122218	Push wheelchairs
11	222712	Clocks and timepieces
12	122434	Devices to protect wheelchairs and their occupants from sunlight or precipitation
13	122439	Devices to which a wheelchair is attached that facilitate movement up and down stairs

Table 6.16: *Feature subset used for predicting a successful programme based on the first two screenings using the logistic regression classifier.*

The forwards subset selection algorithm combined 7 features to ensure the highest AUC for random forest. In table 6.17 on the facing page, the features are listed by importance.

Rank	Feature	Description
1	nTrainingPrWeek	-
2	HasRollator	-
3	MinEvaluation	-
4	Sex	-
5	TimeBetweenTrainingsAvg	-
6	MaxEvaluation	-
7	NeedsEnd	-

Table 6.17: *Feature subset used for predicting a successful programme based on the first two screenings using the random forest classifier.*

Discussion This is the experiment using information from the rehabilitation programme to predict whether the citizens will successfully complete the rehabilitation. As can be seen in the best feature subsets the amount of trainings each week and mean time between trainings are important to the prediction. For the random forest classifier, this is the first experiment where neither the *DevicesUnique* nor *DevicesCount* predictors are used to find the best AUC. As earlier described, this experiment has the largest amount of candidate predictors - 21 in total before the assistive device predictors are added. The model uses the feature *HasRollator*, which is a categorical feature describing whether the citizen has lent a rollator, and this is the second most important feature. The average amount of trainings per week *nTrainingPrWeek* is the most important for the model. However, as can be seen in figure 6.7 on the preceding page, the model is fairly good in predicting true negatives, but not at predicting true positives.

6.1.6 Discussion of model selection

The overall best performing logistic regression was found in the NEEDS-8 experiment with an AUC of 0.827. The best random forest classifier was found in experiment NEEDS-4 with an AUC of 0.878. In all experiments the *DevicesUnique* or the *DevicesCount* predictors were used, except for SP-B, where random forest used the DigiRehab defined AT predictors to ensure the best AUC. The *NeedsStart* predictor occurred in five of the ten subsets. In four of the five occurrences, it was the most important predictor, and in the last occurence it was rated second most important.

Even though experiment SP-A contains the largest observation set, both classifiers performed the worst in terms of AUC. A likely explanation for this is noise in the data, as the feature vector includes all citizens if they have at least two screenings, as opposed to the other feature vectors. This is described in section 4.4.2 on page 39.

In some experiments, it was observed how the logistic regression performed well in predicting the true negatives while random forest yielded good prediction for true positives and vice versa. Therefore, it would be of interest to combine these models in an ensemble model which could improve the prediction accuracy. This could, however, reduce the transparency.

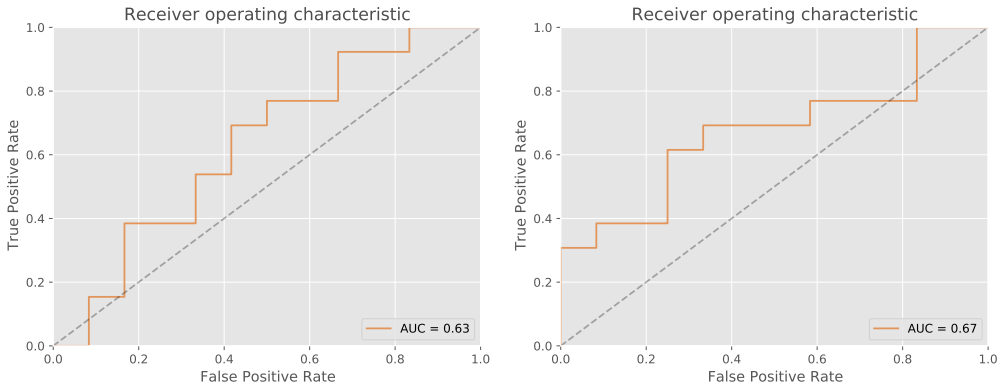
There were no obvious correlation between assistive technology ISO classes between the feature subsets. Looking at the main categories, some were more susceptible to appear than others. Especially category 12 and 22 frequently recurred. These are the assistive products for personal mobility and transportation, and the assistive products for communication and information management. Category 24 (Assistive products for controlling, carrying, moving and handling objects and devices) and 15 (Assistive products for domestic activities and participation in domestic life) only occurred in one experiment each. The model selection yielded the best performing subset for all experiments using forwards subset selection with a maximum of 20 predictors. As a different feature selection approach could yield other combinations of features, it cannot be concluded whether the overall best subsets actually were found.

6.2 Model assessment

The purpose of this section is to present the results of the model assessment. All results in this chapter are obtained based on predicting on the hold-out data sets. The hold-out sets make up 20% of the data and are used to test the models on new data after the subset selection. The number of test samples in the hold-out sets are as follows: feature vector 1: 25, feature vector 2: 70, feature vector 3: 44. The five experiments are described in the succeeding sections where the ROC curves of each experiment are presented. These provide insights to the relationship between the *false positive rate* and the *true positive rate*. At the end of this section the assessment results are summarised in table 6.18 and discussed and compared to the model selection results.

6.2.1 Experiment NEEDS-0: Predicting an improvement in *need for help* score by at least 0

The logistic regression classifier with the highest AUC from model selection is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.8a. The random forest classifier with the highest AUC from model selection is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.8b. The highest AUC in this experiment is obtained by the random forest classifier with 0.67, while the AUC of logistic regression is 0.63.



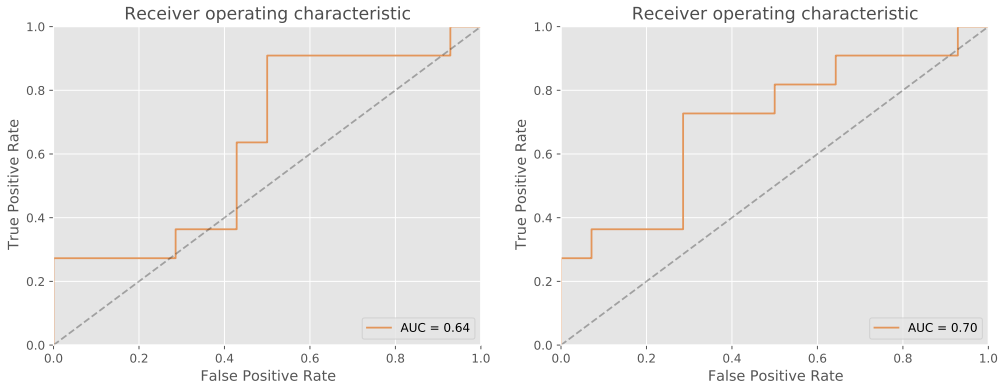
(a) *Logistic regression.* **(b)** *Random forest*

Figure 6.8: *ROC curves for predicting an equal or decreased need for help score for logistic regression (a) and random forest (b).*

6.2.2 Experiment NEEDS-4: Predicting an improvement in *need for help* score by at least 4

The logistic regression classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.9a on the next page. The random forest classifier with the highest AUC is used to predict on the hold-out data and to plot

the ROC curve shown in figure 6.9b. The highest AUC in this experiment is obtained by the random forest classifier with 0.70, while the AUC of logistic regression is 0.64.



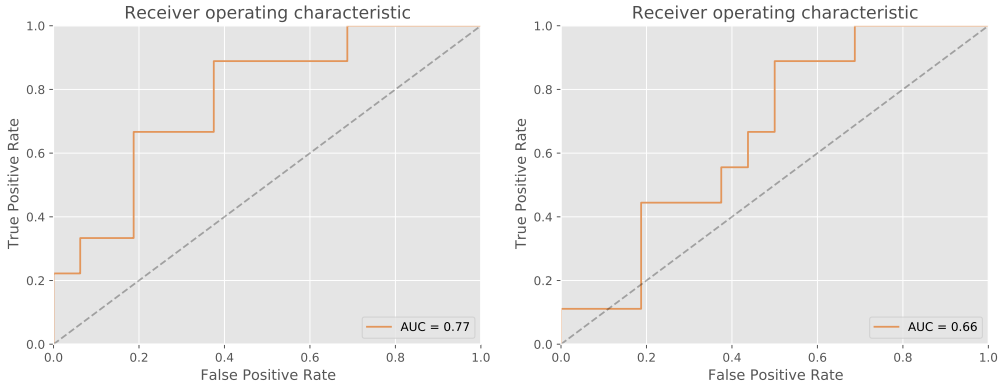
(a) *Logistic regression.* **(b)** *Random forest*

Figure 6.9: *ROC curves for predicting a decrease in the need for help score by at least 4 for logistic regression (a) and random forest (b).*

6.2.3 Experiment NEEDS-8: Predicting an improvement in *need for help* score by at least 8

The logistic regression classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.10a.

The random forest classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.10b. The highest AUC in this experiment is obtained by the logistic regresssion classifier with 0.77, while the AUC of random forest is 0.66.

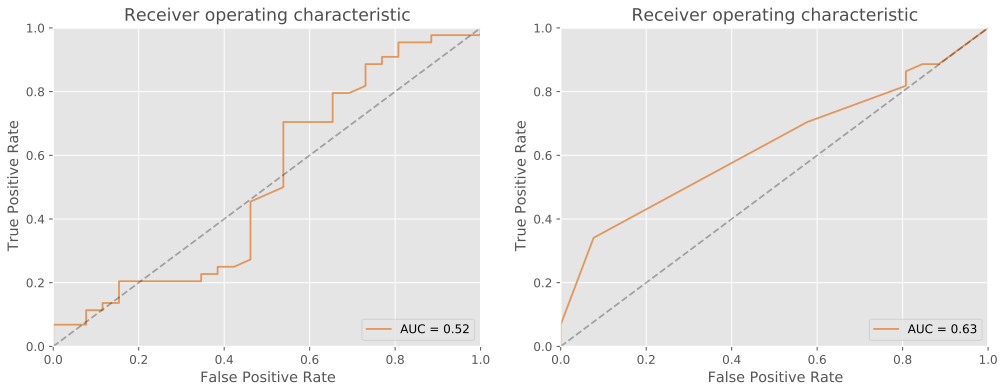


(a) *Logistic regression.* **(b)** *Random forest*

Figure 6.10: *ROC curves for predicting a decrease in the need for help score by at least 8 for logistic regression (a) and random forest (b).*

6.2.4 Experiment SP-A: Predicting whether the citizen completes a rehabilitation programme based on the first screening

The logistic regression classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.11a. The random forest classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.11b. The highest AUC in this experiment is obtained by the random forest classifier with 0.63, while the AUC of logistic regression is 0.52.

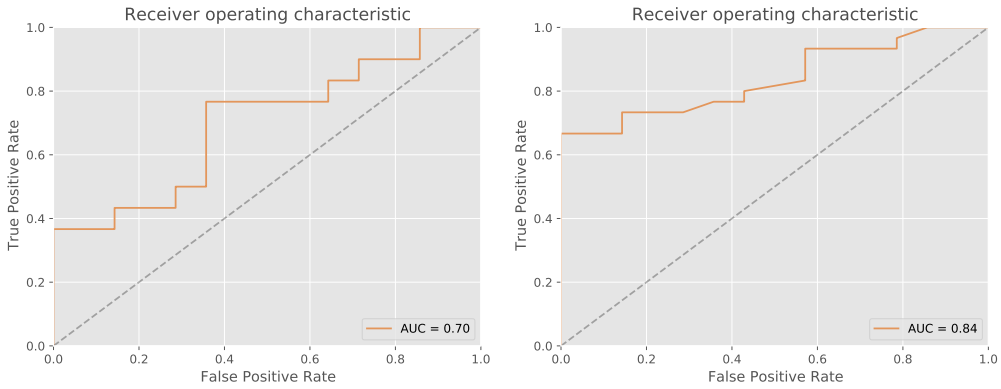


(a) *Logistic regression.* **(b)** *Random forest*

Figure 6.11: *ROC curves for predicting whether the citizen completes a rehabilitation programme based on the first screening using logistic regression (a) and random forest (b).*

6.2.5 Experiment SP-B: Predicting whether the citizen completes a rehabilitation programme based on the first two screenings

The logistic regression classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.12a on the next page. The random forest classifier with the highest AUC is used to predict on the hold-out data and to plot the ROC curve shown in figure 6.12b on the following page. The highest AUC in this experiment is obtained by the random forest classifier with 0.84, while the AUC of logistic regression is 0.70.



(a) *Logistic regression.* **(b)** *Random forest*

Figure 6.12: *ROC curves for predicting whether the citizen completes a rehabilitation programme based on the first two screenings using logistic regression (a) and random forest (b).*

6.2.6 Summary of model assessment

In table 6.18 the results of the model assessment are presented along with the AUCs from the model selection. From this it can be seen how the performance of each model on the hold-out set compares to the AUC obtained at the model selection step. It can be seen that the assessment performance generally is lower than that of the model selection. In all experiments except NEEDS-8, the assessment AUC is higher for random forest than for logistic regression. The highest performing model from the model selection was random forest in the NEEDS-4 experiment with an AUC of 0.878. However, this performance has not upheld in the assessment, where the same model obtained an AUC of 0.70 on the hold-out set. The best performing model in the assessment is random forest in SP-B experiment, which achieved an assessment AUC of 0.84. This was the second best performing model in the selection step. The second best performing model in the assessment is logistic regression for the NEEDS-8 experiment with an assessment AUC of 0.77.

Experiment	Model	AT predictors	AUC	
			Model selection	Model assessment
NEEDS-0	Logistic regression	DevicesUnique	0.732	0.63
	Random forest	DevicesUnique	0.743	0.67
NEEDS-4	Logistic regression	DevicesUnique	0.783	0.64
	Random forest	DevicesUnique	0.878	0.70
NEEDS-8	Logistic regression	DevicesCount	0.827	0.77
	Random forest	DevicesUnique	0.759	0.66
SP-A	Logistic regression	DevicesUnique	0.687	0.52
	Random forest	DevicesUnique	0.654	0.63
SP-B	Logistic regression	DevicesUnique	0.814	0.70
	Random forest	DigiRehab-defined[1].	0.863	0.84

Table 6.18: *Results for assessment of the classifiers. For comparison both the AUC from the model selection and the AUC from the model assessment are included.* [1] *i.e. HasRollator*

6.2.7 Discussion of model assessment

As the results show, all AUCs of the model assessment are lower than the AUCs obtained during model selection. This indicates that inspite of using 5-fold cross-validation during model selection and regularisation for logistic regression the models have all managed to overfit the data. This confirms that having a hold-out set is important to avoid overconfidence in the model based on the model selection results. The best performing model is as mentioned above, the random forest for SP-B. This is also the model with smallest difference in the AUCs. This is interesting as this model is the only one that uses the DigiRehab-defined device feature subset. Training on this subset results in the fewest candidate features. Thus, the number of candidate features may play a role in the overfitting variance. As the other models use *DevicesUnique* or *DevicesCount* as subsets, where the devices are included as dummy features, they have had over 100 candidate features to choose from which may increase the risk of overfitting. The random forest model for SP-B also has the smallest subset of features resulting from the selection. This may also have helped avoid overfitting. The limit on the number of features was set to 20 features for all models as described in section 3.2.5. To avoid overfitting it may have been preferable to have defined a stricter limit on the number of features.

From the result in both selection and assessment it seems that the models have a hard time predicting who will complete a training program prior to any training as the case is for experiment *SP-A*. Even though this data set has more data, than the other experiments this yields the lowest results in both model selection and model assessment with assessment AUCs of 0.52 and 0.63 for logistic regression and random forest, respectively. For logistic regression this is barely above random guessing. However when data from the first four to eight weeks are included the results are much better as seen with the results from *SP-B*.

6.3 LIME explanations

LIME has been used to provide local approximations of predictions. In the following a few examples of the output from LIME are presented a long with some basic information about the subject of the prediction. Both experiments are based on the *SP-B* experiment.

6.3.1 Logistic regression

The logistic regression classifier yielded an AUC of 70% on the hold-out set as described in section 6.2. Two of the citizens in the hold-out set are described below along with the resulting explanation.

6.3.1.1 Citizen A

The majority of predictors for citizen A is shown in table 6.19. She is a woman of 86 years, currently lending a raised toilet seat, a bathtup seat, a rollator with a portable ramp and a gripping tong. Her *need for help* score is 42 and her physical strength is 22. By use of the logistic regression classifier, she is predicted to benefit from a rehabilitation programme with a probability of 71%, which is also the true class label for this citizen. Using LIME, figure 6.13 is presented for the explanation.

Citizen A	
Age	86
NumberATsRunning	12
Sex	Female
RehabIndicator	1.91
NeedsStart	42
PhysicsStart	22
nTrainingPrWeek	1.4
Current ATs (count)	Description
09 12 18 (1)	Raised toilet seats fixed to toilet
09 33 05 (1)	Bathtub seats
12 06 06 (1)	Rollators
18 30 15 (1)	Portable ramps
24 21 03 (1)	Manual gripping tongs
True label	positive
Prediction	positive

Table 6.19: *Citizen A information. (Note that ATs is short for Assistive Technologies).*

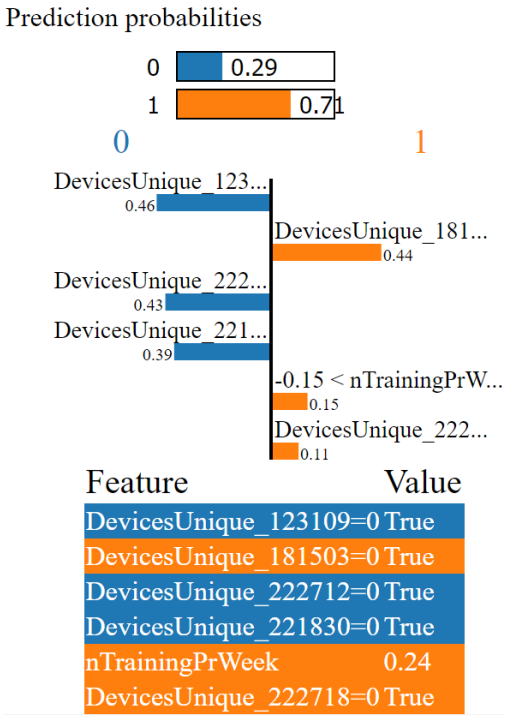


Figure 6.13: *Citizen A LIME explanation*

As can be seen, LIME provides a graphical overview of the explanation. Starting from the top, the prediction probabilities for the citizen is displayed. These are derived directly from the model and states that the citizen will complete a successful programme with a likelihood of 71%. The next part is derived from the explanation and is therfore an approximation based on the neighborhood of the prediction. It shows the six features, that are the most important for this citizen to be classified as able to complete a programme. A list of the features with full names are listed below the graph. For citizen A it is seen that if she were to have one of the devices *123109* (Not mounted rails for self-lifting), *222712* (Clocks and timepieces) or *221830* (Telecoil amplifiers), she would have probably been more prone to belong to the other class. The fact that she does not have device *181503* (Leg extenders) is important for her to be within the category she is assigned to. As the features are standardised for logistic regression, the *nTrainingsPrWeek* does not equal the correct value and the explanation is of less use. The actual list of most important features for logistic regression can be found in table 6.16 on page 72. The most important feature is *nTrainingsPrWeek* followed by *181503* (Leg extenders). The fifth and sixth most important are *123109* (Not mounted rails for self-lifting) and *221830* (Telecoil amplifiers). LIME actually succeeded in finding important features by local approximation.

6.3.1.2 Citizen B

The majority of predictors for citizen B is shown in table 6.20 on the next page. He is 77 years old and is currently in possession of a shower chair, a rollator, a chair and an emergency alarm system. His *need for help* score is 3 and his physical strength is 50. By use of the logistic regression classifier, he is predicted to not benefit from a rehabilitation programme with a probability of 85%, which is also the true class label for this citizen. Using LIME, figure 6.14 on the following page is presented for the explanation.

LIME has again gathered the six most important features for the citizen - and they do not correspond with the six assigned to citizen A. From the explanation it can be derived that him having an alarm system actually points towards him not being able to complete a rehabilitation programme. His number of trainings per week also draws him towards the prediction. The actual list of most important features for logistic regression can be found in table 6.16 on page 72. The most important feature is *nTrainingsPrWeek* followed by *181503* (Leg extenders). For this citizen, Leg extenders is deemed to be most important. Some of the other features approximated by LIME are comparable to citizen A. These are *221830* (Telecoil amplifiers), *123109* (Not mounted rails for self-lifting) and *222712* (Clocks and timepieces).

Citizen B	
Age	77
NumberATsRunning	5
Sex	Male
RehabIndicator	0.0600
NeedsStart	3
PhysicsStart	50
nTrainingsPrWeek	0.8
Current ATs (count)	Description
09 33 07 (1)	Shower chairs with and without wheels
12 06 06 (1)	Rollators
18 09 03 (1)	Chairs
22 27 18 (1)	Personal emergency alarm systems
True label	negative
Prediction	negative

Table 6.20: *Citizen B information. (Note that ATs is short for Assistive Technologies).*

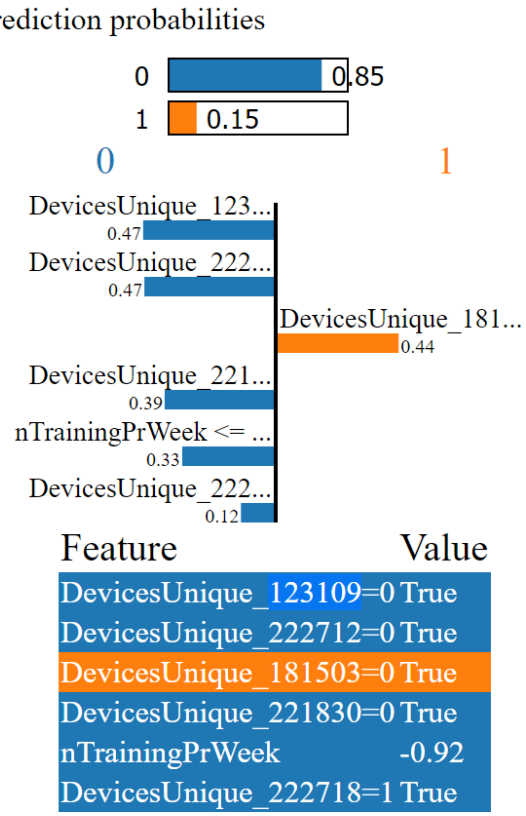


Figure 6.14: *Citizen B LIME explanation*

6.3.1.3 Model intrinsics of logistic regression for comparison

In table 6.21 the coefficients of the logistic regression model are presented to provide a means of comparison.

Feature	Coefficient	Subset selection rank
nTrainingPrWeek	1.04	1
123109	0.78	5
181503	−0.54	2
222712	0.41	11
221830	0.36	6
222718	−0.26	3
TimeBetweenTrainingsAvg	−0.24	7
122439	−0.24	13
122218	0.23	10
NumberATsRunning	0.22	4
122434	−0.21	12
093307	−0.20	8
181226	0.20	9

Table 6.21: *The logistic regression coefficients for prediction of experiment SP-B. Sorted by the highest absolute value of the coefficients.*

6.3.2 Random forest

The random forest classifier yielded an AUC of 84% on the hold-out set as described in section 6.2. Two of the citizens in the hold-out set are described below, along with the resulting explanation.

6.3.2.1 Citizen C

The majority of predictors for citizen C is shown in table 6.22. She is a woman of 75 years, currently lending a raised toilet seat. Her *need for help* score is 58 and her physical strength is 26. By use of the random forest classifier, she is predicted to benefit from a rehabilitation programme with a probability of 73%, which is also the true class label for this citizen. Using LIME, figure 6.15 on the following page is presented for the explanation.

Citizen C	
Age	75
NumberATsRunning	7
Sex	Female
RehabIndicator	2.23
NeedsStart	58
PhysicsStart	26
TimeBetweenTrainingsAvg	4.5
nTrainingPerWeek	1.6
MinEvaluation	2
MaxEvaluation	6
NeedsEnd	55.0
HasRollator	False
Current ATs (count)	Description
09 12 18 (1)	Raised toilet seats fixed to toilet
True label	positive
Prediction	positive

Table 6.22: *Citizen C information. (Note that ATs is short for Assistive Technologies).*

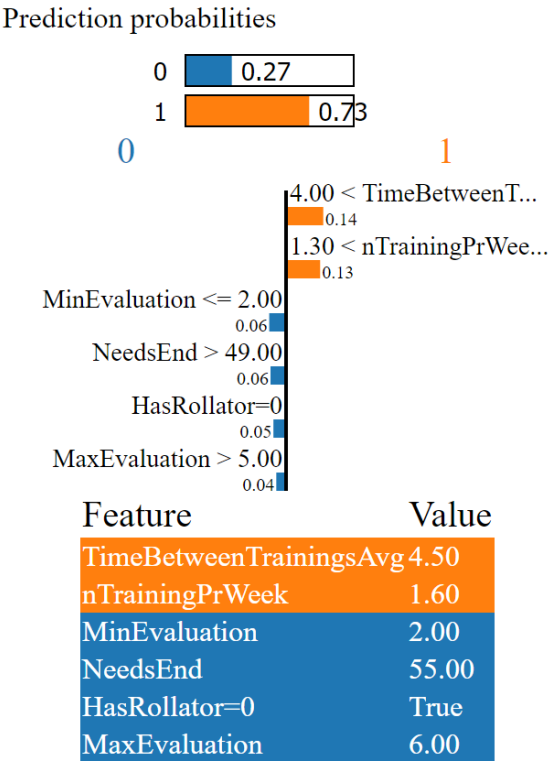


Figure 6.15: *Citizen C LIME explanation*

The explanation shows that due to her *TimeBetweenTrainingsAvg* is above 4.0, this points toward the prediction of her completing a rehabilitation programme. That *nTrainingPrWeeksAvg* is above 1.3 supports this decision as well. This predictor is also considered the most important by the random forest model as seen in table 6.17 on page 73. It is of interest that it uses the *NeedsEnd* predictor. This is the *need for help* score at the time of the second screening and it can be seen that she already had succeeded in reducing the score in the interval. However, LIME states that a score above 50 points towards her not being able to complete a rehabilitation programme. The feature could have been structured in another way, to show the development instead of the actual value to accommodate this.

6.3.2.2 Citizen D

The majority of predictors for citizen D is shown in table 6.23. He is 56 years old and is currently in possession of a rollator and some accessory to this. His *need for help* score is 43 and his physical strength is 31. By use of the random forest classifier, he is predicted to benefit from a rehabilitation programme with a probability of 54%. This is not the true class label for this citizen. Using LIME, figure 6.16 on the facing page is presented for the explanation.

Citizen D	
Age	56
NumberATsRunning	8
Sex	Male
RehabIndicator	1.39
NeedsStart	43
PhysicsStart	31
TimeBetweenTrainingsAvg	5.75
nTrainingPrWeek	1.2
NeedsEnd	67
MinEvaluation	3
Current ATs (count)	**Description**
12 06 06 (1)	Rollators
12 07 24 (1)	Accessories attached to assistive products for walking to hold or carry objects
True label	negative
Prediction	positive

Table 6.23: *Citizen D information. (Note that ATs is short for Assistive Technologies).*

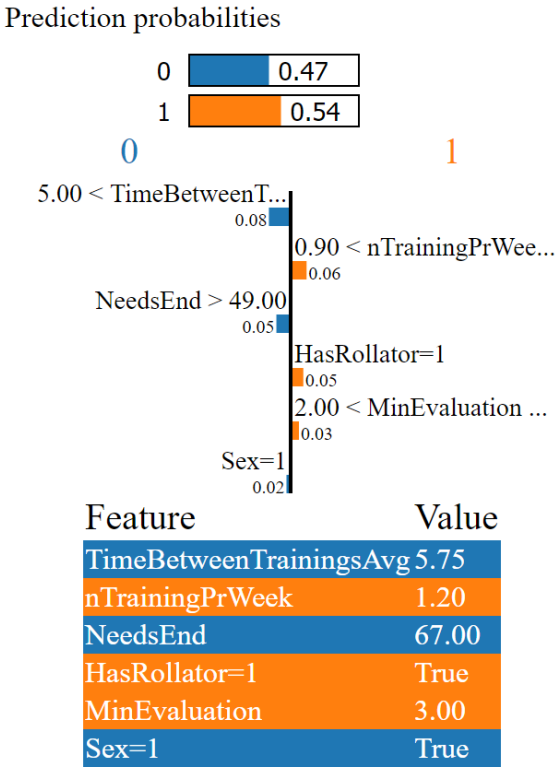


Figure 6.16: *Citizen D LIME explanation*

From the graph it can be seen that LIME had a hard time selecting important features for this citizen. It is also of interest to observe that the *NeedsEnd* threshold differs from the one applicable for citizen C even though the model is the same. This shows that the approximation is only local.

6.3.2.3 Model intrinsics of random forest for comparison

In table 6.24 the feature importance from the logistic regression model are presented to provide a means of comparison.

Feature	Coefficient	Subset selection rank
nTrainingPrWeek	0.29	1
TimeBetweenTrainingsAvg	0.27	5
NeedsEnd	0.22	7
MaxEvaluation	0.093	6
MinEvaluation	0.063	3
HasRollator	0.039	2
Sex	0.031	4

Table 6.24: *The random forest feature ranking for prediction of experiment SP-B. Sorted by the highest value of the feature importances.*

6.3.3 Discussion

From the four citizens, LIME was able to find local approximations that in most cases could substantiate the prediction by the model. It provided a clear overview of how the predictors affected the prediction which aided in getting an insight of the models. Predicting for a standardised data set for the logistic regression made the explanation more unclear, considering that it was harder to obtain an understanding of the thresholds that LIME placed. For the random forest prediction it was more lucid.

Logistic regression has intrinsically transparent properties, as the coefficients of the function provides insights to feature importance. The actual coefficients are depicted in table 6.21 on page 83

As the features are standardised, the coefficients provide possibility of directly comparing the feature importances, which is beneficial compared to if the features varied more in the values. So standardising the data improves global transparency of the relationship between features, wheres it reduces the transparency in the local explanations provided by LIME. Furthermore, the coefficients also provide information on whether each feature affects the outcome in a positive or negative direction as seen by the sign of the number.

For the random forest model it is possible to derive the feature importances. This yields a set of numbers presented in table 6.24 As seen from this table, three features are specifically important. However this does not disclose information on how the features affects the classification as further described in 3.2.3.

7

Discussion

7.1 Comparison to state of the art

It is a challenge to compare the results of the present thesis to previously reported models for predicting the benefit of rehabilitation in home care, as this thesis focuses specifically on the physiotherapy-based rehabilitation rather than rehabilitation in a broader term, which is the focus of the related work in the field of home care. Furthermore, the data sets are different and different methods are applied. However, in [20], Zhu, Chen, Hirdes, *et al.* have trained a k-nearest neighbour and later the same team trained a support vector machine model in [19], both for predicting rehabilitation potential for home care patients. 24,724 Canadian home care citizen were included in the studies. As the evaluation metrics differ, a direct comparison of performance is a challenge. However, the study also defines the most important variables in the predictions. This is done by the use of support vectors far from the decision boundary in the support vector machine model. They define three of which two are related to motivation and prospect, which are not available in the present thesis. The third is related to the bathing capabilities of the citizen. For some of the experiments in this thesis, the assistive technology predictors related to bathing (shower chairs and bath boards) are included in 7 out of 10 feature subsets. This could support the confidence in bathing capabilities as a good predictor of rehabilitation potential.

In [13] Lin, Chen, Tseng, *et al.* have developed models to predict rehabilitation potential in post stroke patients. This is a different application as the subjects are admitted to in-hospital rehabilitation programme after suffering from an acute condition. However, the methods used are comparable as logistic regression, random forest and support vector machine classifiers were applied to predict rehabilitation potential. The performance of the models in terms of AUC were: *logistic regression: 0.755, random forest: 0.769, and support vector machines: 0.777*. These are obtained using 5-fold cross validation on a data set of 313 patients. This means that no hold-out set has been used. In the present thesis the models were evaluated on a hold-out set and overfitting resulting from the feature selection process was seen inspite of using 5-fold cross-validation. Thus, this should be considered when comparing to [13], which does not use a hold-out set. However, with only 20 candidate predictors, this may be less prone to overfitting. Nonetheless, when comparing the assessment AUC of the present thesis with the results achived by Lin, Chen, Tseng, *et al.*, they generally achieved a slightly higher AUC.

The problem researched by Lin, Chen, Tseng, *et al.* was a three class problem while this thesis investigated a binary classification problem. More patients were available for their research, while no model assessment was performed, making it difficult to directly compare the results.

Another study of predictive decision support in clinical settings is conducted by Horng, Sontag, Halpern, *et al.*, who achieved an AUC of 0.86 on a data set of 230,936 patients when including free text in their model. This could indicate a potential for improvement in predictive performance if from more data and leveraging free text comments from practitioners. In the DigiRehab system free text comments exist, but for confidentiality reasons it was not possible to include them in the present study.

7.2 Data and collaboration

The present thesis has been made in collaboration with Aalborg Municipality and DigiRehab. This entailed that the data used in the project is from real world settings where variations and irregularities are found. Thus, a big part of this thesis has been to investigate how such data can be utilised in order to leverage the information within. A lot of focus has concerned the utilisation of data about assistive technologies and how this can be used as predictors. Furthermore, no subjects were filtered from the data based on illness or falls. Thus, subject who did experience set backs as a result of falls or illness may be the cause of noise in the results, but at the same time this avoids introducing bias.

The data sets used in the development of the models were all quite small with 125, 219, and 350 subjects. Originally it was the plan that more data should be made available. However, this was not achieved. This would have included DigiRehab data from Viborg Municipality in which case the model might have suffered less from overfitting and have been able to generalise more to new data and thus, have yielded a higher assessment AUC.

Furthermore data about falls was also not obtained as first planned.

A benefit in the collaboration has been the possibility of leveraging domain knowledge when designing the solution. This has been used to establish one of the representations of assistive technology predictors. This representation was also selected as the most useful representation in the experiment SP-B of random forest, which achieved the highest accuracy on the assessment data. Furthermore, the definitions of *benefit* in the introduction, were also defined with input from DigiRehab.

7.3 Results

The performance of the developed models varies across the different experiments and there is a clear tendency to lower AUC values in the model assessment results compared to the model selection results. A possible reason for this may the high dimensionality of the data set, resulting from the assistive devices being encoded as dummy

variables. This results in a high number of candidate predictors. As the number of observations is rather low, this might lead to overfitting even though cross-validation was used and the maximum number of parameters was constrained to 20. Restricting the number of allowed features further, and gathering more data to increase the number of observations might also increase predictive performance. Alternatively, dimensionality reduction methods could be applied to reduce the number of features. An example of this is *principle component analysis*, where the new features are created based on the greatest variance in the data [41, pp. 230-233]. However, this greatly reduces the transparency of the model as it is no longer clear which of the original predictors contributed to a prediction, but it could be investigated how this could be utilised in combination with LIME to improve predictive performance while still providing transparency at the level of a local approximation.

7.3.1 Leveraging the information of assistive technology

An objective of this thesis is to investigate how assistive technology can be utilised to provide information useful to the predictions, and when looking into the results of the subset selection it is clear that all models yield higher AUCs when including the assistive device features compared to not including them. This indicates that the assistive technology does in fact provide useful information for the models. The choice of device feature representation varies between the three possible representations. While the most commonly selected are the dummy variable-based representations, these also seem to have a higher tendency to result in overfitting, whereas the devices selected by utilising the domain knowlegde of DigiRehab seem to result in less overfitting for SP-B where they are included. However, as the experiments are different, this should be evaluated for the individual experiments.

For the dummy variable-based representations of the device features it was investigated to what level of detail the ISO-class numbers of the devices should be used and a length of six digits was selected. It could be interesting to investigate whether reducing the level of detail can increase model assessment performance as this would reduce the total number of devices, which may help reduce overfitting.

7.3.2 Predictive performance

Firstly, looking at the different NEEDS experiments, which investigate the reduction in the *need for help* score, the assessment AUC ranges from 0.63 to 0.77. These are all well above *random*, and as hyper parameter tuning has not been applied for the classifiers, this provides a good foundation for further improvement in a hyper parameter search.

When looking at the two *successful programme* experiments, some interesting results arise. While SP-B for random forest achieves the highest assessment AUC of all the experiments, SP-A yields the lowest scores on both logistic regression and random forest and for logistic regression this is barely above random classification. This is seen even though SP-A uses feature vector 2, which has the largest number of observations. This may indicate that the predictors available prior to any training are less useful for

predicting whether a citizen will complete a training programme. The features available in SP-B contains information about how a citizen has started out, and this seems to provide good indicators as to whether the programme will be successful. Another consideration is the difference in the data available to the two models. As they are trained using different feature vectors with different subjects this may influence the predictive performance. Using feature vector 2 SP-A has more subjects which cancel the programme than SP-B, which uses feature vector 3, where only citizens that stay in the rehabilitation programme for at least four weeks are included. This difference may lead to a more noisy data set for SP-A. Citizens may drop out of the programme early for a number of different reasons. Examples of patterns that are hard to predict include citizens who are actually making progress during the rehabilitation programme, but then experience falls or other accidents that force them to drop out of the programme. It could be interesting to investigate whether SP-A would yield a better performance if it was trained and assessed using the same subjects as SP-B. This might provide insights to the extent of how the performance difference is related to noise in the data set or a lack of useful predictors.

From the model selection results of the SP-B experiment the confusion matrices showed that the random forest classifier correctly predicted 90.0% of the true negative samples, whereas the prediction of the true positives was considerably worse at 43.2%. However, logistic regression performed well in both cases and employing an ensemble model could therefore be of interest to increase the prediction capability. This could, however, come with a cost of a less transparent model.

7.3.3 Feature selection

Another interesting result relating to SP-A and SP-B is that for these, the *sex* is included in the feature subsets for three out of the four models, whereas it is only included in one of the six NEEDS subset results. This may indicate that whether a citizen can reduce their need for home care help may not be as related to the sex of the citizen, but when predicting who will complete a training programme there is a difference pertained to the sex of the citizen. This is important to pay attention to and consider how to handle this information. When providing clinical decision support it is important to ensure that it is unbiased to provide equal access to health care. Thus, it is relevant to investigate why being male or female seems to be an indicator of whether a citizen will complete a successful programme, and possibly adjust procedures to retain both genders in the programme as both genders seem equally likely to benefit in terms of reducing their *need for help* score if they stay in the programme for 12 weeks.

When looking into the assistive device predictors there do not seem to be a clear pattern as to which devices are selected as part of the best features subsets. A total of seven different categories were in play and within the categories the occurrence of the specific six-digit assistive devices also differed a lot. One reason for this could be pertained to correlation between the devices such that one device may act as a surrogate of another in a different model. Thus, the correlation between device features could be of interest to investigate further. Another possible cause may be the high dimensionality

of the data set, resulting from the assistive devices being encoded as dummy variables. The resulting high number of candidate predictors might make it hard for the models to find patterns.

7.3.4 Transparency

LIME did in many of the cases find a set of the most important predictors, which it provided a local approximation for. In general, the explanation was intuitive and provided an insight concerning the features and how they affected the prediction outcome.

Compared to the intrinsic coefficients which could be directly derived from the trained logistic regression model, LIME provided a illustrative explanation which could be easily read, even though the approximated features in some cases differed from the actual features. This is another trade-off. When applying transparency to models, it is highly important to state under which circumstances the explanations are derived. Thus, one should know whether the explanation is based on an approximation or the actual values to ensure that the explanation can be correctly understood and trusted. A survey among health care professionals could be of use to gain a further insight into how the explanations could be adopted and whether they succeed in increasing trust in a clinical decision support system.

8

Conclusion

Community-dwelling citizens receiving home care experience a gradual decline in physical capacity and receive rehabilitation to counter this. However, as described in the introduction, the referral process differs across municipalities and uncertainties in the clinical judgement of rehabilitation referrals pose a challenge. Especially as it is important to prioritize scarce resources of the health care sector. In the light of these challenges, the prospect of potentially using available data to drive the development of decision support in the area will be a major benefit to the life quality of citizens receiving home care. Furthermore it would profit the health care system in reduced cost and a more appropriate use of resources.

This thesis has broadly examined a data-driven approach to predicting physiotherapy-based rehabilitation benefit of citizens in home care. The work was conducted in collaboration with Aalborg Municipality and DigiRehab and provides a baseline for the further studies on the KL signature project *Intelligent rehabilitation and targeted public assistance for citizens*.

An extensive work of preparing the raw data was conducted in order to fit the project scope. Classification models based on scientifically proven machine learning algorithms have been designed, implemented and tested. This is done while carefully ensuring transparency through a focus on choosing the classifications algorithms to apply, and by using state of the art explainable AI. The models have been applied to data provided by the collaborators and thorough feature extraction, selection and the use of robust validation methods have yielded prediction AUCs similar to comparable studies.

A careful examination of the objective of determining whether a citizen will benefit from physiotherapy based rehabilitation resulted in the definition of two sub-objectives. These support a broad approach to estimating a citizen's rehabilitation potential and how information regarding loans of assistive technology can be utilised to improve the prediction. In all experiments, the use of assistive technology information improved on the prediction performance.

The models were optimised in terms of feature selection to increase prediction performance. Prediction capabilities were finally assessed on an independent validation set to ensure a trustworthy measurement for the model's generalisation performance. With the use of LIME, illustrative explanations for the prediction of the citizen's rehabilitation potential were created which were matched with the intrinsic models to

investigate their strengths and drawbacks. This provided a great insight into the model and its usage of the predictors.

8.1 Contributions

This section outlines the most significant contributions presented in this thesis.

- This thesis defined two distinct and objective definitions for evaluating whether a citizen benefits from a rehabilitation programme, applicable for the signature project *Intelligent rehabilitation and targeted public assistance for citizens*.
- An extensive work of preparing the raw data was conducted in order to fit both objectives.
- The work shows that information about assistive technology can provide useful information to improve the predictive performance of classification models to determine the benefit of physiotherapy-based rehabilitation in home care. No published work was found that study this.
- The representation of the assistive technology was examined as well as the granularity of the ISO classes. It was seen that optimising the length of the ISO class numbers could improve predictive performance. On average a two-level ISO class number granularity yielded lower AUCs than longer ISO class numbers.
- Two models representing two different levels of transparency were compared to evaluate on the prediction capabilities on the data set. In 4 out of 5 times, random forest performed better in terms of the model assessment AUC.
- It was shown that a local model-agnostic explainer (LIME) could be utilized to provide intuitive explanations for a specific prediction outcome and thereby increase transparency of the decision support system.
- A proven python-based framework for carefully examining and transforming rehabilitation data has been established in a manner that allows for state of the art classification. This provides a solid foundation for the continuing work of the signature project.

8.2 Personal outcome

Throughout this project we have greatly expanded our knowledge within decision support systems. This includes the importance of carrying out a large scaled project with raw data sets which had to be carefully examined and processed in order to facilitate the task of classification.

The work was conducted in collaboration with external partners. This has provided us with the experience of working on large-scaled projects with externally defined constraints and objectives closely related to a specific domain. Furthermore, at the initiation of this thesis, there were issues regarding data processing and confidentiality. This had to be handled carefully to ensure privacy of the citizens, and applying the best solution in collaboration with the project partners were especially rewarding. Further information regarding this is found in appendix **??**.

Moreover, experience within comprehending complex research articles within different domain-specific areas was obtained. A vast number of these articles were within the field of machine learning, but medical research also presented a large amount of the articles, which demanded and strengthened our knowledge within this field. Especially within home care, rehabilitation and decision support systems for predicting various diseases.

The ethical aspect of this project also played an important role as we have been working with humans and the complex task of examining who should be offered a rehabilitation programme. In this context, there have been a lot of reflection and consideration. Both with regards to the interest of the individual, for which we sought to predict for - but also with regards to not overstepping the line to other domains, i.e. medical staff.

8.3 Future work

In the present thesis the results achieved using state of the art methods in the development of data-driven decision support provides a sound basis for further exploration and optimisation of relevance to the signature project and the Danish healthcare system. Suggestions for future work are discussed in this section.

Firstly, the focus regarding performance optimisation of the classifiers has in the present thesis been related to feature engineering, where the potential in the provided data has been investigated and useful features have been extracted. For future work it will be relevant to investigate how tuning parameters in model training can further optimise the developed models. For random forest some important hyper parameters include the number of estimators, the maximum features considered for splitting a node, and the maximum depth of a decision tree. For logistic regression the optimisation function and the method of regularisation are interesting hyper parameters to explore further.

Additionally it has been discussed how the models might benefit from more data. As more municipalities are getting involved in the project data from these should be included with the aim of improving the generalisation of the models.

Furthermore, as mentioned in chapter 4, the DigiRehab platform provides care givers with the opportunity of adding comments in a text field. Due to not having a Data Processing Agreement, these descriptions were omitted from the data. If access to this was to be granted, it would be interesting to determine the possible benefit of including this as predictor in the models. In a study of predicting sepsis for in-

hospital patients, utilizing comments from nurses improved the performance of the applied models [11].

Another aspect of the data worth investigating, is how a cluster representation of the assitive technology can leverage the history and sequence of devices to determine the rehabilitation potential of a citizen. However, care should be taken to ensure that it is created with transparency in mind.

Finally it could be useful to study how methods for time series data can be applied to develop a method that tracks the development of a citizen through the duration of the rehabilitation programme to provide decision support throughout the programme. However, the transparency of such methods must be considered.

A

Fall risk factors and prediction. A preliminary analysis

A.1 Fall risk factors

Ageing eventually causes frailty which includes reduction of physical fitness, loss of muscle strength, worse coordination and reduced balance [2]. The handling of the functional impairment is usually a steadily increasing amount of home care combined with usage of assistive devices to compensate the physical decline. However, this treatment usually results in a downward spiral where the citizens depend on multiple assistive devices while their frailty rises [3]. A study conducted in 2018 found higher prevalence of falls among a group of elderly citizens using assistive devices compared to non-users [4]. The falls for users of assistive devices usually happened at a time where they did not use their device. Multiple studies show that being physically active diminishes the effects of frailty and increases self-reliance among elderly [5]. The exercise must be customised to the individual and must be supervised in order to follow the progress and adjust the exercises. It is shown that not all citizens benefit from training and therefore the selection of citizens susceptible to the training program must be made carefully.

The literature identifies above 400 separate risk factors for falls [67] that roughly can be divided into two categories: modifiable and non-modifiable. The non-modifiable risk factors include age, gender, earlier falls and surrounding factors such as the flooring, illumination etc. The modifiable factors include chronic diseases that affect the balance and motor coordination as well as unsuitable glasses and usage of certain medications [68].

A.2 Fall prediction

Most efforts in research regarding fall prevention target clients that have already fallen[12, 67, 69, 70], whereas only few look into trying to predict first-time falling.

By use of decision tree analysis a resent study [12] has developed an algorithm to assess the risk of first time falling for home care clients in Canada. First time falls are in the study defined as falls where the clients has not fallen in the past 90 days. From the

resulting decision tree they identified risk clusters producing 6 categories of fall risk, the lowest being 5-10 % risk of falling and the highest being 31-35 % risk of falling. The tree-model was trained on data from 126,703 home care clients in Canada based on the Resident Assessment Instrument-Home Care (RAI-HC) which includes information such as symptoms, function, and quality of life.

B

Description of features

Feature name	Description
CitizenId	Anonymized identifier for citizens provided by KMD.
PatientId	Anonymized identifier for citizens provided by DigiRehab.
Sex	Gender of the citizen.
BirthYear	Birthyear of the citizen.
Age	Age of the citizen.
StartDate	Date for first screening.
EndDate	Date for last screening in the selected interval.
nWeeks	The amount of weeks between StartData and EndDate.
MeanEvaluation	The mean evaluation score for all trainings in the interval.
StdEvaluation	The standard deviation for all evaluation scores in the interval.
MinEvaluation	The minimum evaluation score for all trainings in the interval.
MaxEvaluation	The maximum evaluation score for all trainings in the interval.
nTraining	The count of trainings done by the citizens in the interval.
nTrainingOptimal	The optimum number of trainings is defined as 2 times per week.
nTrainingPrWeek	The average number of trainings per week.
nTrainingPrWeekMax	The amount of trainings in the week with the most trainings.
nTrainingPrWeekMin	The amount of trainings in the week with the least trainings.
nWeeksWithTrainings	The amount of weeks in the interval with trainings.
nWeeksWithoutTrainings	The amount of weeks in the interval without trainings
TimeBetweenTrainingsAvg	The average amount of days between trainings.
nCancellations	The amount of trainings the citizen has cancelled.
TimeBetweenCancelAvg	The average amount of days between cancelled trainings.
TimeBetweenCancelMax	The maximum amount of days between cancellations.
TimeBetweenCancelMin	The minimum amount of days between cancellations.
nCancellationsPerWeekAvg	The average amount of cancelled trainings in a week.
nCancellationsPerWeekMax	The maximum amount of cancelled trainings in a week.
nCancellationsPerWeekMin	The minimum amount of cancelled trainings in a week.
NeedsStart	The need for help score at the first screening.
NeedsEnd	The need for help score at the last screening.
NeedsDiff	The difference between NeedsStart and NeedsEnd.
NeedsReason	The reason for a worsening of the need for help score.

PhysicsStart	The physical strength score for the first screening.
PhysicsEnd	The physical strength score for the last screening.
PhysicsDiff	The difference between PhysicsStart and PhysicsEnd.
PhysicsReason	The reason for a worsening of the physical strength score.
RehabIndicator	The result of NeedsStart divided by PhysicsStart.
Exercises	A list of exercises for the citizen's trainings.
NumberATsRunning	The amount of assistive products the citizen currently has.
NewAts	The new assistive products lent to the citizen within the interval.
LastStatusDate	Date for when the citizen last changed its status.
LastStatus	The citizen's current status.
HasRollator	True if the citizen currently possesses a rollator.
HasRaisedToiletSeat	True if the citizen currently possesses a raised toilet seat.
HasShowerStool	True if the citizen currently possesses a shower stool.
HasRaisedToiletSeat-AndShowerStool	True if the citizen currently possesses a raised toilet seat (cont.) and a shower stool
DevicesCount	A list of all device categories the citizen currently possesses.
DevicesUnique	A list of all unique device categories the citizen currently possesses.
Cluster	The assistive devices cluster that the citizen belongs to.

Table B.1: *Description of features.*

Clusters for the assistive technologies

Feature vector 1		Feature vector 2		Feature vector 3	
Cluster	Count	Cluster	Count	Cluster	Count
5	23	5	58	5	42
2	13	14	32	14	26
14	11	9	28	16	17
16	10	2	26	3	17
4	8	3	26	9	16
9	8	16	24	2	14
3	7	4	20	4	13
1	6	1	12	29	8
7	4	7	11	12	8
6	3	12	11	7	7
24	3	29	10	24	7
20	3	10	10	1	6
15	3	0	8	8	5
13	2	24	7	22	5
12	2	6	6	10	4
0	2	8	6	11	4
33	2	15	6	34	4
22	2	19	6	0	4
25	2	22	5	19	4
18	1	20	5	6	3
11	1	34	5	13	2
10	1	13	4	33	2
8	1	11	4	18	2
23	1	18	4	20	2
26	1	33	3	15	2
28	1	25	3	21	1
29	1	26	2	25	1
30	1	30	2	26	1
32	1	31	2	28	1

34	1	23	1	31	1
		21	1	32	1
		28	1		
		32	1		

Table C.1: *The count of citizens assigned to one of the 36 clusters for the three different set of subjects associated with the experiments. The clusters are sorted based on the number of citizens assigned to each cluster.*

List of Figures

List of Tables

References

[1] Danmarks Statistik. (2019). Andelen af ældre i pleje- og ældreboliger falder, [Online]. Available: https://www.dst.dk/da/Statistik/nyt/NytHtml?cid=28347.

[2] T. Strandberg, K. Pitkälä, and R. Tilvis, "Frailty in older people", *European geriatric medicine*, vol. 2, no. 6, pp. 344–355, 2011.

[3] G. Häggblom-Kronlöf and U. Sonn, "Use of assistive devices – a reality full of contradictions in elderly persons' everyday life", *Disability and Rehabilitation: Assistive Technology*, vol. 2, no. 6, pp. 335–345, 2007. DOI: 10.1080/17483100701701672.

[4] A. de Oliveira Cruz, S. M. M. Santana, C. M. Costa, L. V. G. da Costa, and D. D. Ferraz, "Prevalence of falls in frail elderly users of ambulatory assistive devices: A comparative study", *Disability and Rehabilitation: Assistive Technology*, vol. 0, no. 0, pp. 1–5, 2019, PMID: 30907182. DOI: 10.1080/17483107.2019.1587016. [Online]. Available: https://doi.org/10.1080/17483107.2019.1587016.

[5] B. Pedersen, "Which type of exercise keeps you young?", English, *Current Opinion in Clinical Nutrition and Metabolic Care*, vol. 22, no. 2, pp. 167–173, Mar. 2019, ISSN: 1363-1950. DOI: 10.1097/MCO.0000000000000546.

[6] T. D. of Civil Affairs. (2019). Serviceloven § 83 a. Accessed: 2020-02-14, [Online]. Available: https://www.retsinformation.dk/Forms/R0710.aspx?id=209925#id94b75b98-941a-4464-958d-7947cee0b658.

[7] H. Lauritzen, M. Bjerre, L. Graff, T. Rostgaard, F. Casier, T. Fridberg, "Rehabilitering på ældreområdet. Afprøvning af en model for rehabiliteringsforløb i to kommuner", 2017.

[8] B. U. of the Ministry of Social Affairs and the Interior, *Rehabilitering på ældreområdet efter § 83a i serviceloven*. 2019.

[9] C. Cunningham, F. Horgan, and D. O'neill, "Clinical assessment of rehabilitation potential of the older patient: A pilot study", *Clinical rehabilitation*, vol. 14, no. 2, pp. 205–207, 2000.

[10] G. Valdes, C. B. Simone II, J. Chen, A. Lin, S. S. Yom, A. J. Pattison, C. M. Carpenter, and T. D. Solberg, "Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making", *Radiotherapy and Oncology*, vol. 125, no. 3, pp. 392–397, 2017.

[11] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, and L. A. Nathanson, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning", *PloS one*, vol. 12, no. 4, 2017.

[12] A. Kuspinar, J. P. Hirdes, K. Berg, C. McArthur, and J. N. Morris, "Development and validation of an algorithm to assess risk of first-time falling among home care clients", *BMC geriatrics*, vol. 19, no. 1, p. 264, 2019.

[13] W.-Y. Lin, C.-H. Chen, Y.-J. Tseng, Y.-T. Tsai, C.-Y. Chang, H.-Y. Wang, and C.-K. Chen, "Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation", *International journal of medical informatics*, vol. 111, pp. 159–164, 2018.

[14] T. Larsen, L. Mark, S. Cichosz, P. Secher, and O. Hejlesen, "Population exacerbation incidence contains predictive information of acute exacerbations in patients with chronic obstructive pulmonary disease in telecare", English, *International Journal of Medical Informatics*, vol. 111, pp. 72–76, 2018, ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2017.12.026.

[15] M. H. Jensen, S. L. Cichosz, B. Dinesen, and O. K. Hejlesen, "Moving prediction of exacerbation in chronic obstructive pulmonary disease for patients in telecare", *Journal of telemedicine and telecare*, vol. 18, no. 2, pp. 99–103, 2012.

[16] C. Ngufor, H. Van Houten, B. S. Caffo, N. D. Shah, and R. G. McCoy, "Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin a1c", *Journal of biomedical informatics*, vol. 89, pp. 56–67, 2019.

[17] L. Cheng, M. Zhu, J. W. Poss, J. P. Hirdes, C. Glenny, and P. Stolee, "Opinion versus practice regarding the use of rehabilitation services in home care: An investigation using machine learning algorithms", *BMC medical informatics and decision making*, vol. 15, no. 1, p. 80, 2015.

[18] H.-F. Mao, L.-H. Chang, A. Y.-J. Tsai, W.-N. Huang, and J. Wang, "Developing a referral protocol for community-based occupational therapy services in taiwan: A logistic regression analysis", *PloS one*, vol. 11, no. 2, 2016.

[19] M. Zhu, Z. Zhang, J. P. Hirdes, and P. Stolee, "Using machine learning algorithms to guide rehabilitation planning for home care clients", *BMC medical informatics and decision making*, vol. 7, no. 1, p. 41, 2007.

[20] M. Zhu, W. Chen, J. P. Hirdes, and P. Stolee, "The k-nearest neighbor algorithm predicted rehabilitation potential better than current clinical assessment protocol", *Journal of clinical epidemiology*, vol. 60, no. 10, pp. 1015–1021, 2007.

[21] D. (for Digitisation). (2019). Kommunernes arbejde med kunstig intelligens, [Online]. Available: https://digst.dk/strategier/kunstig-intelligens/signaturprojekter/.

[22] K. L. (G. Denmark). (2019). Signaturprojekter om kunstig intelligens i kommuner og regioner, [Online]. Available: https://www.kl.dk/okonomi-og-administration/digitalisering-og-teknologi/kommunernes-arbejde-med-kunstig-intelligens/.

[23] M. Harbo, "Effektanalyse, Ballerup", 2019.

[24] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, and R. B. Haynes, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review", *Jama*, vol. 293, no. 10, pp. 1223–1238, 2005.

[25] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success", *Bmj*, vol. 330, no. 7494, p. 765, 2005.

[26] G. H. S. R. Group. (Date Unknown). Inforehab home care. Accessed: 2020-02-21, [Online]. Available: `https://uwaterloo.ca/geriatric-health-systems-research-group/research/inforehab/inforehab-projects/inforehab-home-care`.

[27] F. B. Larsen, M. H. Pedersen, K. Friis, C. Glümer, and M. Lasgaard, "A latent class analysis of multimorbidity and the relationship to socio-demographic factors and health-related quality of life. a national population-based study of 162,283 danish adults", *PloS one*, vol. 12, no. 1, 2017.

[28] R. Denmark, *Pres på sundhedsvæsenet*, 2015. [Online]. Available: `https://www.regioner.dk/media/2209/2015-pres-paa-sundhedsvaesenet.pdf`.

[29] T. J. Loftus, P. J. Tighe, A. C. Filiberto, P. A. Efron, S. C. Brakenridge, A. M. Mohr, P. Rashidi, G. R. Upchurch, and A. Bihorac, "Artificial intelligence and surgical decision-making", *JAMA surgery*, 2019.

[30] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care", *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.

[31] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine", *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[32] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission", in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.

[33] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"", *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[34] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes", *Philosophy & Technology*, vol. 31, no. 4, pp. 611–627, 2018.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?" explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[36] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface", in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 2390–2395.

[37] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[38] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. [Calster], "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models", *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, 2019, ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2019.02.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895435618310813.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York Inc., 2009.

[40] L. Breiman, "Random forests", English, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324.

[41] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014, ISBN: 1461471370.

[42] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation", in *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, ser. COLING-02, USA: Association for Computational Linguistics, 2002, pp. 1–7. DOI: 10.3115/1118853.1118871. [Online]. Available: https://doi.org/10.3115/1118853.1118871.

[43] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization", *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[44] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization", *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[45] scikit-learn developers. (). Sklearn.ensemble.randomforestclassifier, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[46] E. Lewinson. (). Explaining feature importance by example of a random forest, [Online]. Available: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e.

[47] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization", *Journal of machine learning research*, vol. 13, no. Feb, pp. 281–305, 2012.

[48] T. Hastie, R. Tibshirani, and R. J. Tibshirani, *Extended comparisons of best subset selection, forward stepwise selection, and the lasso*, 2017.

[49] T. Fawcett, "An introduction to roc analysis", *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[50] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.", *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[51] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms", *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[52] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement abstract", *SIGKDD Explorations*, vol. 12, pp. 49–57, Jan. 2010.

[53] D. Chicco, "Ten quick tips for machine learning in computational biology", *BioData mining*, vol. 10, no. 1, p. 35, 2017.

[54] P. Domingos, "A few useful things to know about machine learning", *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[55] Magasinet Pleje, "Aalborg oplever fald i behov for hjemmehjælp", *Magasinet Pleje*, p. 7, Nov. 2016. [Online]. Available: `https://digirehab.dk/wp-content/uploads/2018/12/Aalborg-magasinetpleje.pdf`.

[56] The National Board of Social Services. (2020). Assistdata, [Online]. Available: `https://hmi-basen.dk/en/`.

[57] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls", *Bmj*, vol. 338, b2393, 2009.

[58] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging", *Cell*, vol. 153, no. 6, pp. 1194–1217, 2013, ISSN: 0092-8674. DOI: `https://doi.org/10.1016/j.cell.2013.05.039`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0092867413006454`.

[59] D. D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions : A review", 2014.

[60] J. Gravgaard, H. D. Macedo, C. F. Pedersen, *(DUBBAH) First analysis of the data: A preliminary report*, `https://dubbah.github.io/`, 2019.

[61] V. E. Kristensen. (). Data om brug af hjælpemidler skal løfte kommunal service, [Online]. Available: `https://kommunalsundhed.dk/kunstig-intelligens-skal-loefte-kommunal-service/`.

[62] JetBrains. (). Pycharm, the python ide for professional developers, [Online]. Available: `https://www.jetbrains.com/pycharm/`.

[63] A. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016, ISBN: 9781449369897. [Online]. Available: `https://books.google.dk/books?id=vbQlDQAAQBAJ`.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[65] Pandas. (). Pandas, python data analysis library, [Online]. Available: https://pandas.pydata.org/.

[66] marcotcr. (). Lime: Explaining the predictions of any machine learning classifier, [Online]. Available: https://github.com/marcotcr/lime.

[67] A. Oakley, M. F. Dawson, J. Holland, S. Arnold, C. Cryer, Y. Doyle, J. Rice, C. Hodgson, A. Sowden, T. Sheldon, *et al.*, "Preventing falls and subsequent injury in older people.", *Quality in Health Care*, vol. 5, no. 4, p. 243, 1996.

[68] L. Evron. (2017). Derfor falder de ældre, [Online]. Available: https://dsr.dk/sygeplejersken/arkiv/ff-nr-2017-4/derfor-falder-de-aeldre.

[69] G. Kojima, T. Masud, D. Kendrick, R. Morris, S. Gawler, J. Treml, and S. Iliffe, "Does the timed up and go test predict future falls among british community-dwelling older people? prospective cohort study nested within a randomised controlled trial", *BMC geriatrics*, vol. 15, no. 1, p. 38, 2015.

[70] D. Schoene, S. M.-S. Wu, A. S. Mikolaizak, J. C. Menant, S. T. Smith, K. Delbaere, and S. R. Lord, "Discriminative ability and predictive validity of the timed up and go test in identifying older people who fall: Systematic review and meta-analysis", *Journal of the American Geriatrics Society*, vol. 61, no. 2, pp. 202–208, 2013.